International Journal of Interactive Multimedia and Artificial Intelligence

June 2021, Vol. VI, Number 6
ISSN: 1989-1660

unir LA UNIVERSIDAD EN INTERNET

*Ignorance more frequently begets confidence than does knowledge: it is those who know little, and not those who know much, who so positively assert that this or that problem will never be solved by science.*

*Charles Darwin*

INTERNATIONAL JOURNAL OF INTERACTIVE MULTIMEDIA AND ARTIFICIAL INTELLIGENCE

ISSN: 1989-1660 –VOL. 6, NUMBER 6

## EDITORIAL TEAM

**OPEN ACCESS JOURNAL**

**ISSN: 1989-1660**

The International Journal of Interactive Multimedia and Artificial Intelligence is covered in Clarivate Analytics services and products. Specifically, this publication is indexed and abstracted in: *Science Citation Index Expanded, Journal Citation Reports/ Science Edition, Current Contents®/Engineering Computing and Technology.*

# Editor's Note

The International Journal of Interactive Multimedia and Artificial Intelligence – IJIMAI (ISSN 1989-1660) provides an interdisciplinary forum in which scientists and professionals can share their research results and report new advances on Artificial Intelligence (AI) tools or tools that use AI with interactive multimedia techniques.

The present volume, June volume, consists of 24 articles of diverse applications of great impact in different fields, always having as a common element the use of artificial intelligence techniques or mathematical models with an artificial intelligence base. As is logical, COVID is present in several manuscripts of this volume, always focused on the prediction and estimation of the presence of the disease. In addition to this expected presence, there are manuscripts of a semantic or syntactic analysis nature as well as works in the field of management and recommender systems. It is also worth mentioning several works in the field of video compression and signal processing. Of course, the Internet of Things and text analysis for several applications could not be missed in this volume. Finally, different manuscripts on usability and satisfaction, investments, solar panels, malware detection, video analysis, audio analysis and learning can also be found in this volume.

Volume begins with the most important topic of the present time, COVID-19. Thus, Prada et al. propose a model for mortality risk prediction whose input is the key aspect of COVID, X-ray images. Their approach is based on convolutional neural networks reinforcing learning with patient aspects such as age and gender. This results in better accuracy than previous models. Following the theme of the previous work, Khattak et al. propose an analogous prediction model based on the input of X-ray images and machine learning and deep learning models as in the previous case. In this case, a model called Multilayer Spatial Covid Convolutional Neural Network is proposed, obtaining a success rate in COVID detection of 98%, thus improving previous analogous proposals.

Switching topics, but without leaving the medical field, the volume continues with an article proposed by Singh et al. whose focus is based on feature extraction using deep learning and machine learning models for arrhythmia classification. Thus, by means of classical techniques such as SVM and LSTM, hit rates very close to 99.5% are obtained in the case of SVMs. It is interesting to analyze the statistical approach to feature extraction presented in this manuscript.

Within the same medical subject, the following article proposed by Hassan et al. presents a medical image segmentation model. Several works focus on problems of classification and estimation of certain features, but in this case, the approach is an earlier step, so that it emphasizes image segmentation, a key aspect in many applications in their early stages such as cancer characterization. Thus, in this manuscript a comparative study of the different existing techniques is carried out in order to obtain conclusions about the suitability of some models or others.

Closing the medical theme, and with a view to all potential users of AI models, García-Peñalvo et al. present the CARTIER-IA platform, which brings artificial intelligence algorithms to non-specialized personnel. One of the objectives of the platform is to provide a usable and user-friendly environment so that algorithms can be applied to image-type data, for example.

Jumping from the medical field to the management field, Gil et al. present a complete review of the capabilities of Machine Learning algorithms in project management. This manuscript has more than 150 references that show the amount of work that exists in the literature taking advantage of the capacity of this type of models for application in project management.

The business world also includes recommender systems, as this is one of their main applications. Thus, the following manuscript presents a tool based on one of the most widely used multi-criteria decision techniques, PROMETHEE, for the development of an industrial maintenance application. The power of this work presented by Nawal et al. lies in the use of unsupervised models such as the cluster, which allow knowledge to be extracted where there is none a priori, obtaining an accuracy of 90%, a high value for the case of unsupervised models.

However, if we go into the world of software usability, what are the causes that play a crucial role? This is the question posed by Otten et al. in the following manuscript, who carry out a comparative study using two studies that shed light on the answers to this complex question, which is so important for the world of software engineering.

Linking the last two articles related to recommender systems and usability, Bobadilla et al. present a deep learning model capable of predicting fairness in recommender systems. In this work, the authors rely on an initial knowledge of the users' demographic information.

And Amazon can also be considered as a recommender system, so the following authors, Kumar et al., propose to work on this platform and propose a predictive system for Amazon product reviews. They propose a Machine Learning model able to rate products by analyzing the text of the different reviews, combining a Bayesian and SVM approach. Investment recommendations are those proposed by Martín et al. in the following manuscript, improving on the more classical approach by introducing dynamic selection mechanisms for the optimal decision rule.

One of the most popular topics at present is the sentimental analysis and emotions, so this issue could not miss a manuscript on this subject. Huddar et al. present a model based on bidirectional LSTMs and tested on contrasted datasets in the literature, improving on the most widely used current models. RNNs are presented to capture the state of the interlocutor in order to estimate his or her sentiment.

Continuing with the model presented in the previous article, the RNN, the next manuscript in this volume is presented by Dhanith et al. and propose an analogous model but in this case applied to the detection of words embedded in the web. Thus, a new method is proposed integrating Adagrad optimized Skip Gram Negative Sampling and RNN.

COVID, sentimental analysis and text analysis are presented in this volume, but Internet of Things could not be left out. Thus, Meana-Llorián et al. in their work present a model that aims to integrate Smart Objects within traditional social networks in such a way that allows the connection between people and objects through them, for example, an object can perform an action based on a post on Twitter. This is undoubtedly a new approach with a long way to go.

After the medical field, one of the most sought-after fields for Artificial Intelligence models is the field of engineering. Thus, Rezk et al. present a model based on particle swarm to solve the optimization problem in solar panels. The presented model perfectly balances the two main qualities of a social adaptative algorithm, exploration and exploitation, thus obtaining optimal results.

Within engineering, we can find the new topic of 5G. Here we find the work presented by Gupta et al. where they propose a classifier based on an architecture composed of different models such as

DOI: 10.9781/ijimai.2021.05.008

SVM and boosted trees. In this way, a model capable of predicting propagation loss, an important parameter in network planning, is built.

Within the family of social adaptation algorithms, such as the one presented in the previous works based on particle swarm, is the bee colony algorithm. The following manuscript proposed by Shareduwan et al. presents this model combined with RBF to demonstrate its power against classical metaheuristic models. The results obtained support the proposal put forward.

Changing to the world of software, and focusing on the mobile world, the next manuscript proposed by Dhalaria et al. presents a model for detecting malware in Android systems in order to carry out a classification process. Thus, a hybrid approach is proposed that integrates the analyzed characteristics, thus obtaining good results.

We are in the digital age of data generation and consumption, and in most cases, data is generated via video feeds. Therefore, Ebadi et al. present a new approach to video data compression, which is undoubtedly a problem within a software platform due to the high computational requirement. An iterative approach based on classical spline and least squares theories but applied within the video space is presented.

Given the topic of image-type data, what role does virtual reality play? This question is answered in the manuscript presented by Galán et al. by proposing a rigorous comparative study within virtual reality. The results show that the way images are presented influences the user's perception.

Previous manuscripts have presented approaches for the processing of video signals, but the following manuscript presented by Arronte et al. proposes the analysis of audio data. Specifically, a model based on LSTM and CNN for the classification of the motivational pattern of a song is presented. Thus, a model with an architecture that combines the two methodologies, CNN for feature extraction and LSTM for exploiting the song sequencing, is proposed.

One of the topics of the IJIMAI magazine is based on learning. Thus, Tlili et al. present a smart collaborative educational game for teaching English vocabulary using learning analytics. The results presented in an experimental group versus a control group support the new model presented.

Learning is an important topic, but teamwork can be framed within it, and therefore, the following manuscript presented by Conde et al. proposes a study of the effects of individual use of Telegram on the competence of teamwork development. Therefore, the use of this communication vehicle is evaluated, using learning metrics, within a team structure, analyzing its impact with significant results.

Volume finishes with the line of education and the last manuscript presented by Cervantes-Perez et al. proposes a new approach to adaptive navigation control based on the Bayesian approach. This is a new approach as they propose to replace Bloom's taxonomy with Marzano's taxonomy.

Javier Martínez Torres
Managing Editor
University of Vigo

# TABLE OF CONTENTS

# COVID-19 Mortality Risk Prediction Using X-Ray Images

J. Prada[1], Y. Gala[1], A. L. Sierra[2] *

[1] Universidad Autónoma de Madrid, Cantoblanco, Madrid (Spain)
[2] Universidad Complutense de Madrid, Madrid (Spain)

Received 24 May 2020 | Accepted 16 February 2021 | Published 8 April 2021

unir
LA UNIVERSIDAD EN INTERNET

## Abstract

The pandemic caused by coronavirus COVID-19 has already had a massive impact in our societies in terms of health, economy, and social distress. One of the most common symptoms caused by COVID-19 are lung problems like pneumonia, which can be detected using X-ray images. On the other hand, the popularity of Machine Learning models has grown exponentially in recent years and Deep Learning techniques have become the state-of-the-art for image classification tasks and is widely used in the healthcare sector nowadays as support for clinical decisions. This research aims to build a prediction model based on Machine Learning, including Deep Learning, techniques to predict the mortality risk of a particular patient given an X-ray and some basic demographic data. Keeping this in mind, this paper has three goals. First, we use Deep Learning models to predict the mortality risk of a patient based on this patient X-ray images. For this purpose, we apply Convolutional Neural Networks as well as Transfer Learning techniques to mitigate the effect of the reduced amount of COVID19 data available. Second, we propose to combine the prediction of this Convolutional Neural Network with other patient data, like gender and age, as input features of a final Machine Learning model, that will act as second and final layer. This second model layer will aim to improve the goodness of fit and prediction power of our first layer. Finally, and in accordance with the principle of reproducible research, the data used for the experiments is publicly available and we make the implementations developed easily accessible via public repositories. Experiments over a real dataset of COVID-19 patients yield high AUROC values and show our two-layer framework to obtain better results than a single Convolutional Neural Network (CNN) model, achieving close to perfect classification.

## Keywords

Convolution Neural Network, Coronavirus COVID-19, Deep Learning, Machine Learning, Medical Images.

DOI: 10.9781/ijimai.2021.04.001

## I. Introduction

MACHINE Learning (ML) [1], is a branch of Artificial Intelligence whose objective is to build systems that automatically learn from data. The popularity of ML techniques has grown exponentially in recent years and they have been applied to solve a wide variety of problems, such as stock market prediction [2], fraud detection [3], or renewable energy prediction [4], [5].

Although often considered an independent field, Deep Learning (DL) [6], is not less and not more than just another family of Machine Learning models. However, it is a family of models with some extremely relevant properties, such as its high predictive power and its ability to perform end-to-end learning. A specific family of Deep Learning techniques, called Convolutional Neural Networks (CNNs) [7], presents a set of properties highly advantageous for its use in image classification tasks and has in recent years become the state-of-the art for this type of problems.

Image recognition or image classification problems [8], are a set of tasks among the supervised learning [9] branch of ML problems which goal is to correct segment images into a pre-defined set of possible groups or classes. For instance, we may want to classify if an image contains a car, label 1, or not, label 0. Image classifications tasks show up often in the healthcare sector. Some examples of these problems will be Diabetic Retinopathy diagnosis [10], histological analysis [11], or tumor early detection [12].

Taking this into account, the aim and motivation of this research is to apply these techniques to predict the mortality risk of a COVID-19 patient using X-ray images and demographic data of the patient.

We divided our research in two different phases. The first step of this research is to use CNN models to predict the targeted mortality risk using solely X-ray images as input. We will call this model COVID-CheXNet.

Once this COVID-CheXNet model is built, we aim to train a second model, which will act as a second layer, which will use as input the output of our COVID-CheXNet, numeric information regarding characteristics of the X-ray image, and other basic demographic patient data like gender and age of the patient. For this purpose, we tested some of the most popular and powerful Machine Learning models like Neural Nets [13], Support Vector Machines (SVMs) [14] or Extreme Gradient Boosting (XGBoost) [15], together with Logistic Regression and Random Forest [16] models that will act as benchmarks.

* Corresponding author.
E-mail addresses: jesus.prada@estudiante.uam.es (J. Prada), yvonne.gala@estudiante.uam.es (Y. Gala), analusie@ucm.es (A. L. Sierra).

To test the usefulness of this new framework, experiments using a public dataset of COVID-19 X-ray image data collection are carried out. One of the main difficulties to build these models, often found in healthcare real problems, is the reduced amount of X-ray data available right now for COVID-19 patients, even more reduced when we add to this the necessity of knowing if the outcome of that patient was or not an Exitus. Transfer Learning [17] has shown to be a good method to mitigate the negative effects of this lack of data and will be the approach followed in this paper to try to solve this issue.

Theoretical details and code implementations for this two-layer framework, are developed and made publicly available, as well as datasets used in the experiments.

The novelty of our research is mainly due to two factors. The first one is the aim itself, as to our knowledge this is the first study that tries to predict the mortality risk of a COVID-19 patient using ML models based on X-ray images. The other main novelty factor is our proposed two-layer framework that allow us to combine a CNN prediction based on X-ray images with other numerical sources of information like demographical data of the patient, as past research about using X-ray images to make predictions about other lung diseases has focused solely on the use of a single CNN model.

The rest of this paper is organized as follows. In Section II we compare the motivation and limitation of related works. A brief review of prior theoretical background for the main ML models tested, Deep Learning and CNN basic concepts is presented in Section III. Section IV gives an in-depth description of the proposed method, both COVID-CheXNet layer and the final second ML layer, as well as implementation details. In Section V we describe experiments over a real-world public COVID-19 dataset and show the corresponding results. Section VI analyzes the results obtained in these experiments. Finally, the paper ends with the Section VII on conclusions and possible lines of future work.

## II. Related Work

COVID-19 research publications based on the use of ML techniques are still limited, but some works have some common ground with our research.

CNN models have already been shown to achieve good performance when solving the image recognition problem of classifying if a patient have pneumonia or other lung related diseases based on X-ray images [18]. However, the aim here is different to the more specific task we want to tackle in our research, which is to completely focus only on the COVID-19 disease among all lung related health problems.

Convolutional Neural Networks have also been used to diagnose COVID-19 in patients based on X-ray images [19],[20] or CT scans [21]. However, we aim here to go a significant extra step and predict the mortality risk of this patient. We consider this to be much more helpful for clinicians, as when capable of performing an X-ray scan on a patient, clinicians will in most cases also be able to conduct a test for more accurate COVID-19 diagnosis, tests that are moreover getting cheaper and quicker to analyze with the passage of time.

Furthermore, these related studies directly use CNN models that use as input solely X-ray images. We propose here to combine this CNN predictions with a second layer model that also uses as input other numeric data, like demographical data about the patient and characteristics of the image. This is a critical difference as results show this two-layer framework greatly decreases prediction errors compared with the single CNN layer that only uses X-ray images as input.

Novelty of our approach is confirmed in [22], a recent paper that reviews research of AI applied for fighting coronavirus and that heavily mentions the use of DL techniques to diagnose COVID-19 but does not make any reference that points to the existence to this day of a research about the use of ML to predict mortality risk in these patients.

## III. Prior Theoretical Background

### A. Support Vector Machine

The aim of SVM is to obtain the best separating hyperplane possible between two o several different classes. We will focus here on the 2-class or binary problem. In real-world problems, usually finding a hyperplane which separates perfectly the data is not possible. Therefore, defining the slack variables $\xi = (\xi_1, \xi_2 \dots \xi_N)$, one natural way to define this problem will be

$$\begin{aligned} \max_{\beta,\beta_0} \quad & M \\ subject\ to \quad & y_i(x_i^T\beta + \beta_0) \geq M - \xi_i, i = 1,\dots,N \\ & ||\beta|| = 1 \end{aligned} \tag{1}$$

where M is the margin between the training points for class 1 and -1, $\xi_i$ is the absolute value of the amount by which the prediction $f(x_i) = x_i^T\beta + \beta_0$ is on the wrong side of its margin.

Reference [23] shows that this problem is equivalent to the following convex constrained optimization problem

$$\begin{aligned} \max_{\beta,\beta_0} \quad & \frac{1}{2}||\beta||^2 + C\sum_{i=1}^N \xi_i \\ subject\ to \quad & y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i, i = 1,\dots,N \\ & \xi_i \geq 0, i = 1,\dots,N \end{aligned} \tag{2}$$

where the parameter $C$ is often called *cost*. It is easy to see that the hard margin case corresponds to $C = 1$, that leads to $\Sigma\xi_i = 0$, i.e. not a single point on the wrong side of the margin.

The problem solved in practice is the dual formulation derived using Lagrangian techniques [24].

$$\begin{aligned} \max_{\alpha_i} \quad & L_D(\alpha_i) = \sum_{i=1}^N \alpha_i - \frac{1}{2}\sum_{i=1}^N \sum_{j=1}^N \alpha_i\alpha_j y_i y_j x_i^T x_j \\ subject\ to \quad & y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i, i = 1,\dots,N \\ & \xi_i \geq 0, i = 1,\dots,N \\ & \alpha_i \geq 0, i = 1,\dots,N \\ & \mu_i \geq 0, i = 1,\dots,N, \Rightarrow \alpha_i \leq C, i = 1,\dots,N \\ & \beta = \sum_{i=1}^N \alpha_i y_i x_i \\ & \sum_{i=1}^N \alpha_i y_i = 0 \\ & \alpha_i[y_i(x_i^T\beta + \beta_0) - (1 - \xi_i)] = 0. \\ & \mu_i\xi_i = 0, i = 1,\dots,N \Rightarrow (C - \alpha_i)\xi_i = 0, i = 1,\dots,N \end{aligned} \tag{3}$$

With the called Karush-Kuhn-Tucker conditions as restrictions.

Finally, using the kernel trick and a kernel function, $k(x_i; x_j)$, satisfying *Mercer's condition* [25] we can get the following analogous formulation

$$\max_{\alpha_i} \quad L_D = \sum_{i=1}^N \alpha_i - \frac{1}{2}\sum_{i=1}^N \sum_{j=1}^N \alpha_i\alpha_j y_i y_j K(x_i, x_j) \tag{4}$$

that allow us to extend the previous linear version of the SVM problem to a non-linear one.

### B. Extreme Gradient Boosting

Boosting models aim to combine different individual models, usually called weak learners, into a single final more powerful model, commonly called strong learner.

In Boosting, weak learners are of a homogenous nature, i.e., they all come from the same family of models. Normally this family of models are decision trees or combination of them, Random Forest models.

These weak learners are trained in a sequential fashion. The basic idea is that each individual model would be a simple high bias model, like a shallow tree, and the subsequent weak learner will correct its errors, reducing the bias and increasing the goodness of the final model or strong learner.

In computational terms, the sequential nature of the method could be a drawback, but the aim is that this negative factor gets balanced by the fact that each individual weak learner is a basic low variance model and thus fast to train.

In Gradient Boosting models, the strong learner, S, is defined by the following equation

$$S = \sum_{i=1}^{L} c_i I_i \tag{5}$$

where $I_i$ represents each one of the individual weak learners and $c_i$ their corresponding coefficients.

In summary, Gradient Boosting follows this iterative algorithm:

1. Errors, $E$, are initialized with the target value to be predicted. Therefore, the first weak learner will predict the desired label.
2. Another individual model that predicts errors $E$ is trained.
3. The new individual model is added to our final combined model, with $c\_i$ the coefficient that minimizes the global error of the new combined model $S\_k$.
4. The value of the errors $E = E\ (S\_\ (k))$ corresponding the new combined model $S\_k$ is updated.
5. Steps 2-4 are repeated until the model converges or the maximum number of iterations is reached.

Extreme Gradient Boosting (XGBoost) is just an optimized implementation of standard Gradient Boosting models.

### C. Artificial Neural Net

An Artificial Neural Net (ANN) model is made up of a collection of connected units called neurons, where the output of each neuron is computed by some non-linear function, called *activation function*, of the sum of its inputs. Neuron connections have weights, so activations of different neurons can have bigger impact than others. Neurons of one layer connect to neurons of the preceding and following layers. In between the input and output layers are zero or more hidden layers.

Given a training sample and a target to predict, an ANN will compute all the activation functions from the input layer to the output layer, obtaining a final prediction as a result. We call this a *forward* pass.

Once this forward pass has been performed, we need an algorithm to propagate backwards the error from the units in the output layer to the units in preceding layers to update model weights using techniques like gradient descent. This is called the *backward pass*. This algorithm is called backpropagation and is used to optimize ANNs. The goal of backpropagation is to be able to extend gradient descent to all the layers in the network. Backpropagation defines the error associated to a hidden unit as the weighted average of the errors of the units in the adjacent layer. The gradient descent for a layer $j$, with $k$ as the next layer and $i$ as the previous one, will have the following formulation

$$\frac{\partial E_L}{\partial w_{ji}} = \frac{\partial E_L}{\partial s_j}\frac{\partial s_j}{\partial w_{ji}} = \delta_j \frac{\partial s_j}{\partial w_{ji}} \tag{6}$$

where $E_L$ represents the local error, $w_{ji}$ is the weight of the connection from unit $i$ to unit $j$, $s_j = \int w_{ji} z_i$ the sum of the weighted inputs of unit $j$, $z_i$ the output of unit $i$, and $\delta_i$ the generalized error at unit $j$.

This can be shown [26] to be equivalent to

$$\frac{\partial E_L}{\partial w_{ji}} = (\int \delta_k w_{kj}) F'_j(s_j)\, z_i \tag{7}$$

### D. Deep Artificial Neural Net

The concept of Deep Learning has had different interpretations in recent years. Deep learning is often employed simply to refer to a specific subset of Artificial Neural Networks. It is used to name ANNs with many hidden layers. However, the Deep Learning denomination has also been used to refer to any type of Machine Learning model framework which consists of an iterative process of several optimization steps or layers. An example of this is Deep Belief Networks (DBNs) [27], a type of ML models used for unsupervised learning. Another example of Deep Learning structure using models other than Neural Networks can be found in [28].

Nevertheless, it is true that clearly the link between Deep Learning and Deep Artificial Neural Nets is strong and almost ever-present nowadays. Several factors have probably had an impact on this, including the fact that ANNs schema adapts almost perfectly to the concept of DL framework and some of the first groundbreaking advances in DL corresponding to deep ANNs.

In recent years, the popularity of DL models has increased in a spectacular manner, due to the wide availability of powerful computing facilities, advances on the theoretical underpinnings of multilayer perceptrons (MLPs), several improvements on their training procedures and a better understanding of the difficulties related to many layered architectures, like better weight initialization methods and new activation functions such as Rectified Linear Unit (ReLU). To all these factors we can add the appearance of multiple development frameworks such as TensorFlow [29] and Keras [30].

### E. Convolutional Neural Network

In the past, image classification Machine Learning models used raw pixels to classify the images. You can classify dogs for instance based on color histograms and edge detection, i.e. by color and ear shape. This method has been successful but has its limitations, especially when it encounters images with more complex patterns.

Convolutional Neural Networks are a type of neural network model which allows us to extract higher representations from an image. Unlike the classical image recognition where the image features are defined manually as a previous step, CNN takes the image's raw pixel data, trains the model, then extracts the features automatically for better classification.

This type of approach, where expert knowledge to pre-process the image is not needed, is usually known as *end-to-end learning*, and is one of the main reasons behind the recent popularity of these models.

In its most basic version, CNNs are a combination of two type of layers:

- Convolution layer: sweeps a moving window through images and then calculates the filter dot product of the pixel values. This allows convolution to emphasize relevant features.
- Pooling layer: Replaces output of convolution with a summary to reduce data size and processing time. This summary can be for instance the maximum or mean value among a set of several values. This allows pooling to determine features that produce the highest impact and reduces the risk of overfitting.

### F. Transfer Learning

Until recently, conventional ML and Deep Learning algorithms have been traditionally designed to work in isolation. These algorithms are trained to solve a specific task and the models must be rebuilt from scratch once the task changes.

However, it is well-known that humans have an inherent ability to transfer knowledge from one task to another. What we acquire as knowledge while learning about one task, we can utilize in the same

way to solve related tasks. The more related the tasks, the easier it is for humans to transfer our knowledge.

Transfer Learning method tries to apply this same intuition to Deep Learning models, overcoming the isolated learning paradigm and utilizing knowledge acquired for one task to solve related ones.

The idea is that, when trying to solve a task using DL models, instead of training the model from scratch one can reutilize totally or partially other DL models trained to solve similar tasks. For instance, a model built to detect cats, could be reused to detect instead dogs.

There are four main transfer learning approaches, depending on how much reutilization of the previous model is done:

1. Reutilize only the Deep Learning structure, i.e., the configuration and order of the different layers. All the corresponding weights are trained from scratch using the data related to the new task.
2. Reuse the DL structure and use trained weights as initial values. All the weights will be updated using the new data, in a process usually called *fine tuning*.
3. Reuse the DL structure and the weights of some layers, update the rest. You will select a threshold layer, up until this layer all weights will remain fixed, the layer from this point to the output layer will be updated using the new data.
4. Reutilize the DL structure with the same weights. Model weights will not be adapted to the new task and only extra layers added to the base ones will serve to adapt the model to your task. This can only be a valid option when the two problems are similar.

## IV. Proposed Method

This section aims to describe the technical details of the proposed ML framework to solve the task of predicting mortality risk for a COVID-19 patient. Details of the dataset and experiments carried out to test its efficacy are detailed in Section V.

### A. First Layer

As described in Section I, the aim of our first layer is to build a model able to give a mortality risk using as input only X-ray images from COVID-19 patients, which we will call COVID-19 CheXNet. We decided that for this purpose the most suited family of models were CNN models, as they have proved repeatedly to be the best option in image classification tasks like the one in hand.

As stated before, one of the main difficulties when trying to solve our task was the lack of available data. Due to its novelty, there are not many X-ray images publicly available for patients with confirmed COVID-19 diagnosis. Furthermore, this shortage of availability was multiplied by the fact that in our case the target is the outcome, Exitus or no Exitus, of the patient. Datasets with both X-ray images and patient outcome were difficult to find and their volume small.

To deal with this drawback, we applied two methods: First, we make use of transfer learning techniques to take advantage of the knowledge extracted by CNN models from previous research in similar tasks. Second, we also applied data augmentation methods to create new synthetical X-ray images.

We describe our data augmentation approach in Section V.C, so we will focus here on the transfer learning methodology applied. CheXNet model [31] is a Convolutional Neural Network that achieves Radiologist-Level Pneumonia Detection on Chest X-Rays. It has been shown to have a margin of >0.05 AUROC over previous state of the art results and an F1 score of 0.435 (95% CI 0.387, 0.481), higher than the radiologist average of 0.387 (95% CI 0.330, 0.442). This CheXNet model is trained using a Deep Learning structure called Densenet-121, a 121-layer convolutional neural network, the simplest DenseNet among those designed over the ImageNet dataset. The Densenet-121 structure is shown in Fig. 1.

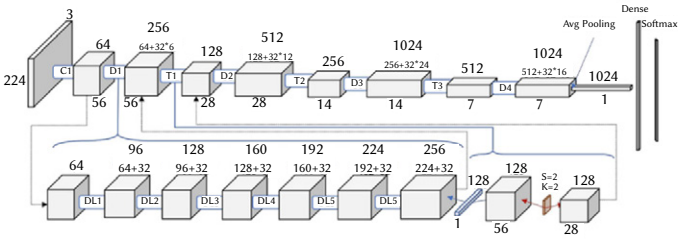



Fig. 1. Densenet-121 layer structure.

We use as our base model this CheXNet CNN. We opted to go for method 4 of transfer learning, as described in Section III.F, i.e., reusing the Densenet-121 structure and preserving the weights of CheXNet, fine-tuning only some additional layer weights to our new COVID-19 dataset.

To the Densenet-121 structure, we added two dense ReLU activation layers with 512 and 256 units, respectively. Finally, we added a logistic layer with sigmoid activation that will generate the final prediction of our model. This is binary classification problem, so only one unit is needed. All these layers are separated by dropout layers.

As our tackled problem represents an example of unbalanced classification task, i.e. there are more cases of non-Exitus label than Exitus outcomes, we set different class weights to balance the impact of each class on the CNN loss function. Therefore, errors in the minority class are penalized more than errors in the majority class.

All weights from the base CheXNet are frozen, i.e. not updated using our new data. Weights from these extra layers will define the correct adaptation of our COVID-19 CheXNet model to the problem we want to tackle.

Implementation of our proposed COVID-19 CheXNet in Python can be found on GitHub[1]. This implementation is based on the use of Keras.

### B. Second Layer

Once we have a mortality risk prediction based solely on X-ray images coming from our first layer CheXNet model, the goal of our second layer model is to combine this prediction output with basic demographic data like gender, age and location, and basic details of the X-ray scan like the view used and the offset, to compute a new mortality prediction. This way we aim to get an improved mortality risk prediction with respect to the one obtained in the first layer, as we are now basing our prediction on additional information.

This is done using the following approach. Mortality risk prediction of layer 1 model becomes the first input column of a new input dataset, that has as remaining columns or input variables information related to demographics and X-ray image characteristics of each patient. As target of this dataset, we will use again the outcome of the patient, Exitus (1) or survival (0). This new combined dataset is passed as input to our second layer ML model to generate new and improved mortality risk predictions as our final output. The total list of input variables used as inputs of this second layer can be found in Table I.

To decide which model to use in this second layer we compute a grid search testing Logistic Regression, Random Forests, SVM, XGBoost and ANN models. The first two are more basic ML families, but we decided to include them due to having a low dimensionality dataset and to at least provide a good benchmark reference.

[1] https://github.com/jesuspradaalonso/COVID-19-CheXNet-

TABLE I. Input Data of Second Layer Model

| Type | Variable |
|---|---|
| **CheXNet** | Mortality risk prediction |
| **Demographics** | Gender |
| | Age |
| | Location |
| **X-ray** | View |
| | Offset |

We carry out hyperparameter optimization for each one of these families of models, as will be described in Section V.D.

The implementation needed to build this second layer model is also available on our GitHub[1], both in R and Python versions.

### C. Two-layer Framework Diagram

Object process diagram of this two-layer framework is presented in Fig. 2.

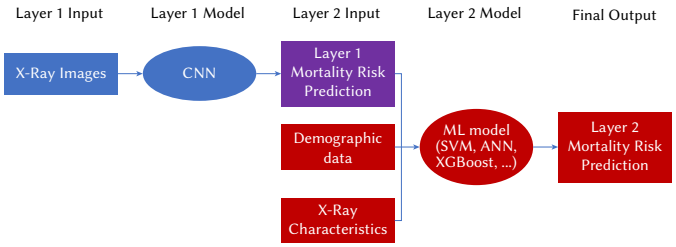



Fig. 2. Proposed two-layer framework diagram.

## V. Experiments and Results

To test the performance of our proposed models described in Section IV we evaluate its goodness over an experiment based on public data available on COVID-19 patients.

### A. Dataset

We used two sources to build our dataset:

- covid-chestxray-dataset[2]: GitHub with information, both X-ray images and basic clinical data, for 209 COVID-19 patients.
- Spanish society of medical radiology, SERAM, COVID-19 data[3]. From this source 12 registers where manually extracted.

Therefore, the combined dataset contains 221 registers. For each register the following information of the patient is available:

- X-ray chest image.
- Gender.
- Age.
- Hospital location.
- X-ray view: anteroposterior (AP) or posteroanterior (PA).
- X-ray offset.

This dataset can be found in our public GitHub repository[1].

### B. Train/val/test Split

Although the optimal ratio of data used in train, validation and test depends on the problem at hand, the most recommended [32] approach is to split the data into 70% for training and 30% for test, and this is the ratio we follow in our experiments.

In our problem the split must be carried out based on patient id, not per row or register. The reason for this is that in the dataset there are some patients with more than one X-ray entry, and it will be a clear case of data leaking to have different images belonging to the same patient in different splits.

This patient-based split has two consequences. First, standard cross-validation implementations, which are row-based, could not be used. Thus, we preferred to use a fixed validation set instead of cross-validation. To create this validation set without reducing more the training set, already small due to data limitations, we decided to use half of the patient ids belonging to the test set as validation.

Second, we applied the 70-30 ratio to the number of rows, the ratio in terms of patient ids used for train and validation/test is different, as not all patients have the same number of X-ray images in the dataset.

Taking all this into consideration, our original dataset is split for training, validation, and test purposes as follows:

1. Train: 65% of patient ids.
2. Validation: 17.5% of patient ids.
3. Test: 17.5% of patient ids.

In addition, the split also considers the class of each case, thus preserving the class imbalance ratio over the three sets of data.

Data augmentation techniques are applied to train and validation sets as explained in the next section.

### C. Data Augmentation

We have already seen that one of the methods to deal with the problem of a small dimensionality in our available dataset is to use transfer learning to reutilize knowledge extracted from other data, as described in Section IV.A

Other popular tool to reduce the impact of this issue is called data augmentation [33]. As having a large dataset is crucial for the performance of the deep learning model, these tools aim to create synthetic examples based on the original dataset.

There are two main approaches to generate these new artificial samples:

- Generate modifications over the original dataset. The changes applied can be of different nature: affine transformations like rotation and translation, perspective transformations, contrast changes, gaussian noise, dropout of regions, hue/saturation changes, cropping/padding, blurring, etc.
- Create images from scratch based on the global distribution found in the original dataset. For this purpose, Generative Adversarial Networks (GANs) [34] are the state-of-the-art.

We decided to apply rotation and contrast modifications for this experiment to create new images, as they are one of the most common changes you can find among real X-ray images carried out in hospitals.

Therefore, if we decide that the batch size used in each epoch when training the CNN model is for instance 32 images, in each epoch of the CNN training process each one of these 32 images would be the result of randomly selecting one of the original training images and then apply random rotation and contrast modifications to it. Thus, we could say that the data pool when using data augmentation consists of an infinite set of images, all of them variations from the original train data pool images.

### D. Hyperparameter Optimization

Each family of Machine Learning models has a set of hyperparameters that are to be optimized to find the optimal model of that family for a given ML task.

Usually this is done by performing a grid search, where you train a different model for each possible combination of hyperparameters you want to analyze, each model is evaluated using a chosen metric over

[2] https://github.com/ieee8023/covid-chestxray-dataset

[3] https://covid19.espacio-seram.com/index.php

a validation set, and the selected hyperparameter values are the ones that correspond to the best performing model. We followed this grid search approach in our experiments.

The detailed list of all the hyperparameters we optimized in our grid search can be found in Table II.

TABLE II. HYPERPARAMETERS OPTIMIZED FOR CHEXNET MODEL USED IN THE FIRST LAYER AND EACH ML FAMILY TRIED AS MODEL IN THE SECOND LAYER

| Model | Hyperparameter |
|---|---|
| **CheXNet** | epochs |
| | batch size |
| | learning rate |
| **Random Forest** | number of trees |
| | nº of candidates at each split |
| | minimum size of terminal nodes |
| **SVM** | cost |
| | gamma |
| **XGBoost** | eta |
| | gamma |
| | max_depth |
| | min_child_weight |
| | subsample |
| | colsample_bytree |
| | num_parallel_tree |
| | nrounds |
| | lambda |
| | alpha |
| **ANN** | number of units |
| | epochs |
| | batch size |
| | learning rate |

### E. Evaluation Metric

As evaluation metric we use the Area Under the Curve (AUC), the most standard evaluation metric for binary classification problems. It is defined as the area under the receiver operating characteristic (ROC) curve, defined by the False Positive Rate (FPR) in the x-axis and the True Positive Rate (TPR) in the y-axis, where:

$$TPR = \frac{TP}{TP+FN} \quad (8)$$

$$FPR = \frac{TN}{TN+FP} \quad (9)$$

where TP, TN, FP, and FN are the true positives, true negatives, false positives, and false negative values, respectively.

### F. Experiment Results

AUC results achieved, for both first and second layer models, are presented in Table III. For the case of the COVID-19-CheXNet model, we also show the difference in performance with or without the use of the data augmentation techniques described in Section V.C.

TABLE III. AUC RESULTS FOR EACH MODEL AND DATASET

| Model | AUC Train | AUC Val | AUC Test |
|---|---|---|---|
| **COVID-19-CheXNet w/o data augmentation** | 0.93 | 0.87 | 0.85 |
| **COVID-19-CheXNet w data augmentation** | 0.93 | 0.93 | 0.94 |
| **Second Layer** | 0.99 | 1 | 1 |

Furthermore, the AUC curve obtained by our COVID-19-CheXNet over the test set is shown in Fig. 3.

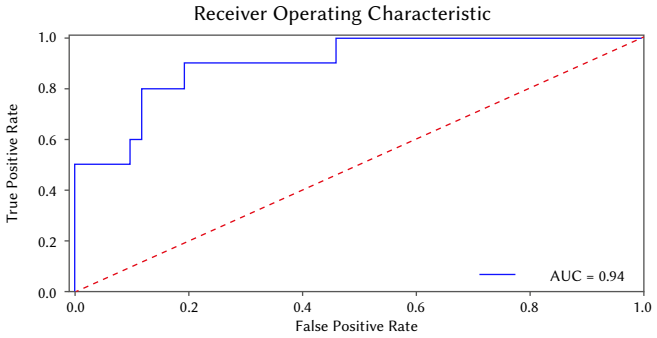



Fig. 3. COVID-19-CheXNet AUC Test. Blue curve represents test AUC for our first layer CNN model predictions. Red dashed line represents a model with an AUC of 0.5 and is used as reference.

We also used heatmaps to visualize which lung areas produced a higher activation in our COVID-19-CheXNet model for deceased patients, which could be useful for practitioner's analysis. One example is shown in Fig. 4.

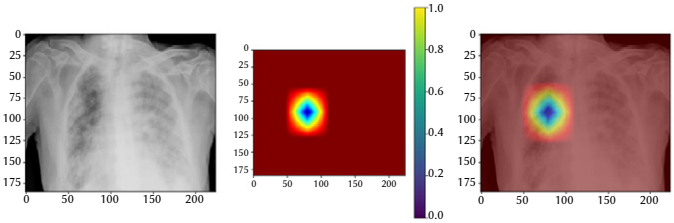


Fig. 4. COVID-19-CheXNet heatmap for a deceased patient.

Finally, we also analyzed the variable importance of each one of the six input variables of the second layer model, by means of conducting a ROC curve analysis on each predictor. Results can be found in Table IV.

TABLE IV. SECOND LAYER MODEL VARIABLES IMPORTANCE IN TERMS OF AUC

| Variable | Rank | AUC |
|---|---|---|
| **pred** | 1 | 0.94 |
| **age** | 2 | 0.71 |
| **sex** | 3 | 0.63 |
| **offset** | 4 | 0.57 |
| **view** | 5 | 0.56 |
| **location** | 6 | 0.53 |

## VI. DISCUSSION

Several conclusions can be drawn from our experiment results shown in Section V.F. First, all our models achieve a high AUC value, above 0.93 for train and 0.85 over test, which seems to point to a good effectiveness of our transfer learning approach, described in Section IV.A.

In addition, the positive impact of using data augmentation is clear comparing the results of COVID-19-CheXNet with and without applying these techniques. 11% and 7% improvement of AUC is achieved over the validation and test sets, respectively. This shows how data augmentation helps the model to generalize better and not suffer from overfitting problems.

Third, our second layer model can achieve close to perfect performance over the test set. Although the exact AUC values obtained could be impacted by the use of a small dataset and the results should be corroborated once larger volumes of COVID-19 X-ray and outcome data are available, the improvement observed between our

first and second layer models performance shows that our intuition that combining mortality risk prediction based solely on X-ray images with other basic demographic and image information could yield even better predictions seems to be valid.

Finally, variable importance analysis shows that the prediction output of the COVID-19-CheXNet first layer model is clearly the factor with greater prediction power among the six predictors used by our second layer model. The top three is completed with age and sex variables, which seems in line with recent research [35] that have already pointed out them as relevant factors in COVID-19 mortality.

The main contributions of this paper are four. First, we aim to predict the mortality risk of COVID-19 patients based on X-ray images to help clinicians lessen the impact of this disease. Some research has been done on the use of Deep Learning models to diagnose COVID-19 based on this type of images, as reviewed in Section II, but we consider that our model predictions can have a bigger positive impact, as diagnosis can always be done using clinical tests once the patients is in the hospital, as would be the case for a patient suitable of getting an X-ray scan.

Second, we propose to add a second layer to this first model using X-ray images, which will use a combination of the prediction of the first layer DL model and basic demographics of the patients and characteristics of the image. This will allow to further optimize final mortality risk predictions, but it is an approach that has received little attention and no approaches like this are found in the literature about COVID-19 prediction models.

Third, we combined two different sources of data to create a unique and novel COVID-19 dataset, providing X-ray images as well as basic demographic information for a total of 221 registers. Data related to COVID-19 is still rare, so we hope this could help further research.

Finally, we make our model implementations and datasets used in our experiment publicly accessible via GitHub, as detailed in Section IV. Principles of Reproducible Research are always recommended but not always followed, and we wanted to be definitive on this aspect.

## VII. Conclusion and Future Work

### A. Conclusions

This paper presents a proposed method to predict mortality risk on COVID-19 patients combining a CNN model based only in X-ray images, with a second layer ML model which uses as input the output of that CNN first layer model together with other basic patient demographic and image technical properties information.

Results show that our proposed method achieves close to or even perfect performance regarding AUC over the test dataset used in our experiments.

Furthermore, results also evidence that our proposed techniques, like transfer learning, data augmentation and the addition of a second layer model improve the overall prediction power of the final model, which seems to confirm out hypothesis and the usefulness of our proposed framework.

### B. Future Work

We know that the main limitation of our research is the small dataset we were obliged to work with due to COVID-19 data availability. Therefore, conclusions drawn from our experiment results should be confirmed with a different and larger dataset. We are currently collaborating with Hm group of hospital in Spain to use a dataset of more than 2310 patients which we hope could greatly enhance our model power and statistical significance of our conclusions. We hope to have experiment results over this new dataset in the coming months.

Furthermore, a more exhaustive optimization of our models in terms of more layer weights being fine-tuned, additional data augmentation techniques being applied, and a bigger hyperparameter grid search being carried out, can be tested to search for a model performance improvement, and we plan to conduct these experiments with the larger dataset earlier mentioned.

Recent proposed frameworks that allow to mix images input with numeric information in a single CNN are suited to the problem we try to tackle. Experiments using these models could be carried out and results compared with our two-layer proposed framework.

Finally, using GANs as data augmentation tool has been shown to improve results obtained by models in healthcare classification tasks [36], and we aim to test it in our proposed framework.

## Acknowledgment

We would like to thank SERAM and the University of Montreal for the public COVID-19 dataset they made available.

## References

[1] C. M. Bishop, *Pattern recognition and machine learning*, New York, USA: Springer, 2006.

[2] T. Kimoto, K. Asakawa, M. Yoda, and M. Takeoka, "Stock market prediction system with modular neural networks", in *International joint conference on neural networks*, San Diego, USA, 1990, pp. 1-6.

[3] S. Maes, K. Tuyls, B. Vanschoenwinkel, and B. Manderick, "Credit card fraud detection using Bayesian and neural networks", in *Proceedings of the 1st international naiso congress on neuro fuzzy technologies*, Havana, Cuba, 2002, pp. 261-270.

[4] Y. Gala, A. Fernandez, J. Diaz, and. J. R. Dorronsoro, "Support vector forecasting of solar radiation values", in *Hybrid Artificial Intelligent Systems*, Salamanca, Spain, 2013, pp. 51-60.

[5] J. Prada, and J. Dorronsoro, "General noise support vector regression with non-constant uncertainty intervals for solar radiation prediction", *Journal of Modern Power Systems and Clean Energy*, vol. 6, no. 2, pp. 268-280, 2018, doi: 10.1007/s40565-018-0397-1.

[6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*, Cambridge, USA: MIT press, 2016.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", in *Advances in neural information processing systems*, Nevada, USA., 2012, pp. 1097-1105.

[8] H. Kaiming, Z. Xiangyu, R. Shaoqing, and S. Jian, "Deep residual learning for image recognition", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, USA., 2016, pp. 770-778.

[9] R Caruana, and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms", *Proceedings of the 23rd international conference on Machine learning*, New York, USA., 2006, pp. 161-168.

[10] R. Gargeya, and T. Leng, "Automated identification of diabetic retinopathy using deep learning", *Ophthalmology*, vol. 124, no. 7, pp. 962-969, 2017, doi: 10.1016/j.ophtha.2017.02.008.

[11] K. Sirinukunwattana, S. Raza, Y. Tsang, D. Snead, I. Cree, and N. Rajpoot, "Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images", *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1196-1206, 2016, doi: 10.1016/ 10.1109/ TMI.2016.2525803.

[12] A. Akselrod-Ballin, L. Karlinsky, S. Alpert, S. Hasoul, R. Ben-Ari and E. Barkan, "A region based convolutional network for tumor detection and classification in breast mammography", Deep learning and data labeling for medical applications, Athens, Greece., 2016, pp. 197-205.

[13] S. Haykin, *Neural networks: a comprehensive foundation*, New Jersey, USA: Prentice Hall PTR, 1994.

[14] C. Chang, and C. Lin, "LIBSVM: A library for support vector machines", ACM transactions on intelligent systems and technology (TIST), vol. 2, no. 3, pp.1-27, 2011.

[15] T. Chen, and. C. Guestrin, "Xgboost: A scalable tree boosting system", in *Proceedings of the 22nd acm sigkdd international conference on knowledge*

*discovery and data mining*, New York, USA, 2015, pp. 785-794.

[16] L. Breiman, "Random forests", *Machine learning*, vol. 45, no. 1, pp. 5-32, 2001.

[17] S. J. Pan, and Q. Yang, "A survey on transfer learning", *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345-1359, 2009, doi: 10.1109/TKDE.2009.191.

[18] J. Zech, M. Badgeley, M. Liu, A. Costa, J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study", *PLoS medicine,* vol. 15, no. 11, pp. 962-969, 2018, doi: 10.1371/journal.pmed.1002683.

[19] I. Apostolopoulos, D. Ioannis, and T. A. Mpesiana, "Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks", *Physical and Engineering Sciences in Medicine*, vol. 1, no. 1, pp.14-26, 2020.

[20] M. Ahsan, T. Alam, T.Theodore and P. Huebner, "Deep MLP-CNN model using mixed-data to distinguish between COVID-19 and Non-COVID-19 patients", *Symmetry*, vol. 12, no. 9, pp.1526, 2020.

[21] S. Ahuja, B.K. Panigrahi, N. Dey, V. Rajinikanth, and T.K. Gandhi, "Deep transfer learning-based automated detection of COVID-19 from lung CT scan slices", TechRxiv, 2020.

[22] S. Fong, N. Dey, and J. Chaki, "AI-enabled technologies that fight the coronavirus outbreak", *Artificial Intelligence for Coronavirus Outbreak*, 2020, pp.23-45.

[23] N. Cristianini, J. Shaew-Taylor, *An introduction to support vector machines and other kernel-based learning methods*, Cambridge, England: Cambridge university press, 2000.

[24] R. Fletcher, *Practical methods of optimization*, New Jersey, USA: John Wiley & Sons, 2013.

[25] H. Q. Minh, P. Niyogi, and. Y. Yao, "Mercer's theorem, feature maps, and smoothing", in *International Conference on Computational Learning Theory, Pittsburgh*, USA, 2006, pp. 154-168.

[26] N. B. Karayiannis, "Reformulated radial basis neural networks trained by gradient descent", *IEEE transactions on neural networks*, vol. 10, no. 3, pp. 657-671, 1999, doi: 10.1109/72.761725.

[27] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations", in *Proceedings of the 26th annual international conference on machine learning*, New York, USA, 2009, pp. 609-616.

[28] D. Díaz-Vico, J. Prada, and J. R. Dorronsoro, "Deep Support Vector Classification and Regression", in *International Work-Conference on the Interplay Between Natural and Artificial Computation*, Almería, Spain, 2019, pp. 33-43.

[29] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, et al., "Tensorflow: A system for large-scale machine learning", in *12th Symposium on Operating Systems Design and Implementation*, Savannah, USA, 2016, pp. 265-283.

[30] A. Gulli, and S. Palm, *Deep learning with Keras*, Birmingham, UK: Packt Publishing Ltd, 2017.

[31] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, et al., "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning", *Computer Vision and Pattern Recognition*, preprint.

[32] M. Stone, "Cross-validatory choice and assessment of statistical predictions", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 36, no. 2, pp.111-133, 1974.

[33] J. Wang, and L. Perez, "The effectiveness of data augmentation in image classification using deep learning", *Convolutional Neural Networks Vis. Recognit*, 2017, preprint.

[34] X. Mao, Q. Li, H. Xie, R. Lau, Z. Wang, and. S. S. Paul, "Least squares generative adversarial networks", in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017, pp. 2794-2802.

[35] J. B. Dowd, L. Andriano, D. M. Brazel, V. Rotondi, P. Block, X. Ding, et al., "Demographic science aids in understanding the spread and fatality rates of COVID-19", *Proceedings of the National Academy of Sciences*, vol. 117, no. 18, pp. 9696-9698, 2020, doi: 10.1073/pnas.2004911117.

[36] E. Wu, K. Wu, D. Cox, and. W. Lotter, "Conditional infilling GANs for data augmentation in mammogram classification", in *Image Analysis for Moving Organ, Breast, and Thoracic Images*, Granada, Spain, 2018, pp. 98-106.

Jesús Prada Alonso

Double Degree in Computer Science and Mathematics at Universidad Autónoma de Madrid, Spain, 2013, double master's in Computational Intelligence and Applied Mathematics, 2015, PhD. in Machine Learning at Universidad Autónoma de Madrid, Spain, 2020. Researcher at Machine Learning Group at Universidad Autónoma de Madrid. Currently working as Machine Learning Manager at Sistemas de Gestión Sanitaria, SIGESA, a Spanish healthcare company, as Machine Learning Specialist at Iberia Express, Spanish airline company, and as professor at Instituto de Empresa, Madrid, and at Escuela de Empresarios, Valencia. Research interests are renewable energy prediction, e-learning, and health, all of them as tasks to solve using Machine Learning or Deep Learning techniques.

Yvonne Gala García

Degree in Mathematics at Universidad Autónoma de Madrid, Spain, 2009, master's in Education, Universidad Autónoma de Madrid, Spain, 2010, master's in Computational Intelligence, Universidad Autónoma de Madrid, Spain, 2012, PhD. in Machine Learning at Universidad Autónoma de Madrid, Spain, currently in progress. Researcher at Machine Learning Group at Universidad Autónoma de Madrid, Spain. Currently working as Machine Learning Manager at Iberia Express, Madrid, and as professor at Escuela de Empresarios, Valencia. Research interests are Machine Learning applied to solar energy prediction, airlines, and health.

Ana Luisa Sierra Bañón

Degree in Biochemistry at Universidad de Navarra, Spain, 2017, master's in Computational Biology, Universidad Politécnica de Madrid, Spain, 2019, master's in Biostatistics, Universidad Complutense de Madrid, Spain, 2019. Currently working as data scientist at Sistemas de Gestión Sanitaria, SIGESA, a Spanish healthcare company. Research interests are computational biology, biostatistics, and Machine Learning applied to health problems.

# Automated Detection of COVID-19 using Chest X-Ray Images and CT Scans through Multilayer-Spatial Convolutional Neural Networks

Muhammad Irfan Khattak[1], Mu'ath Al-Hasan[2], Atif Jan[1], Nasir Saleem[3]*, Elena Verdú[4], Numan Khurshid[5]

[1] Department of Electrical Engineering, University of Engineering & Technology, Peshawar (Pakistan)
[2] College of Engineering, Al Ain University, United Arab Emirates (UAE)
[3] Department of Electrical Engineering, FET, Gomal University, Dera Ismail Khan (Pakistan)
[4] Universidad Internacional de La Rioja, Logroño (Spain)
[5] Smart Earthquake Management (ISD Lab) National Center of Artificial Intelligence UET-Peshawar (Pakistan)

Received 28 July 2020 | Accepted 8 January 2021 | Published 9 April 2021

unir
LA UNIVERSIDAD EN INTERNET

## Abstract

The novel coronavirus-2019 (Covid-19), a contagious disease became a pandemic and has caused overwhelming effects on the human lives and world economy. The detection of the contagious disease is vital to avert further spread and to promptly treat the infected people. The need of automated scientific assisting diagnostic methods to identify Covid-19 in the infected people has increased since less accurate automated diagnostic methods are available. Recent studies based on the radiology imaging suggested that the imaging patterns on X-ray images and Computed Tomography (CT) scans contain leading information about Covid-19 and is considered as a potential automated diagnosis method. Machine learning and deep learning techniques combined with radiology imaging can be helpful for accurate detection of the disease. A deep learning approach based on the multilayer-Spatial Convolutional Neural Network for automatic detection of Covid-19 using chest X-ray images and CT scans is proposed in this paper. The proposed model, named as the Multilayer Spatial Covid Convolutional Neural Network (MSCovCNN), provides an automated accurate diagnostics for Covid-19 detection. The proposed model showed 93.63% detection accuracy and 97.88% AUC (Area Under Curve) for chest x-ray images and 91.44% detection accuracy and 95.92% AUC for chest CT scans, respectively. We have used 5-tiered 2D-CNN frameworks followed by the Artificial Neural Network (ANN) and softmax classifier. In the CNN each convolution layer is followed by an activation function and a Maxpooling layer. The proposed model can be used to assist the radiologists in detecting the Covid-19 and confirming their initial screening.

## Keywords

COVID-19, Machine Learning, Convolutional Neural Network, X-rays Images, CT Scans.

DOI: 10.9781/ijimai.2021.04.002

## I. Introduction

The corona virus infection flared-up in Wuhan, the capital city of Hubei Province, China in December 2019 [1]–[3]. It killed over hundreds and infected more than thousands of people within early few days of the novel corona virus pestilence. The scientists in China named it 2019 novel Corona virus (2019-nCov) [4]. The International Committee of Viruses named it as Severe Acute Respiratory Syndrome Corona Virus-2 (SARS-CoV-2) whereas the infection is named as the Corona virus disease-2019 (Covid-19) [5]-[7]. The subcategories of the corona viruses are alpha-CoV (α), beta-CoV (β), gamma-CoV (γ) and delta-CoV (δ). SARS-CoV-2 is declared a member of the beta-CoV (β) subgroup. People of Kwantung were infected in 2003 by corona virus resulting in Severe Acute Respiratory Syndrome (SARS-CoV). SARS-CoV was also declared to be part of beta-CoV (β) subgroup [8]. SARS-CoV, in 26 countries of Globe, infected over 8000 people with a 9% death rate. Similarly, SARS-CoV-2 infected over 6,728,537 people with a 4% death rate across 202 countries of the World. The infection rate of the SARS-CoV-2 is higher compared to SARS-CoV. The reason for the high infection rate is the regrouping of S Protein in RBD area [9]. Beta-corona viruses infected those people that have close contact with bats [10]-[11]. SARS-CoV-1 and MERS-CoV were transmitted to humans from the cats and Arabian camels. The discovery of the pangolin offspring corona virus and its proximity to SARS-CoV-2 suggested that pangolins can be the possible hosts of the novel 2019 corona viruses [12]. The World Health Organization (WHO) and Centers for Diseases of the US have announced corona virus infection with evidence of human-to-human transfer from five different cases outside China, (Italy [13], US [14], Nepal [15], Germany [16], and Vietnam [17]). On 5 June 2020, SARS-CoV-2 confirmed more than 6,728,537 cases, 3,271,261 recovered cases, and 393,667 death cases. In [18] the statistics about SARS-CoV-2 are shown. Geographical statistics

* Corresponding author.
E-mail address: nasirsaleem@gu.edu.pk

about confirmed Covid-19 cases till June 6, 2020 are obtained from 202 countries (according to the World Health Organization (WHO)). National/International travelling and close contacts with the infected people have been identified as the main reasons of worldwide spread. Huge efforts are being put into developing the vaccines and curing drugs to treat the deadly infection [19]-[20].

Thoracic radiology evaluation is used to diagnose suspected Covid-19 patients. But, scientific methods to identify the virus inside human bodies through machine learning and deep learning using chest x-ray images and computed tomography (CT) scans are potential methods. Timely finding the infection is important in the effort to guarantee the well-timed cure. Machine learning-based studies showed that imaging pattern on the chest x-ray images and CT scans of the patients diagnosed with Covid-19 is a potential analysis tool. The motivation behind the presented study is to detect the Covid-19 using CNN networks. We intend to provide a simple solution with better results. The target of the proposed work is to detect Covid 19 in x-ray images and CT scans efficiently. From the literature it is obvious that performance of relatively simple model VGG 16 is better as compared to the modern GoogleNet and ResNet. Therefore, we are focused on trying a simple version of 2D-CNN inspired from VGG-11. Our CNN network is a subset of VGG-11 which consists of 5 convolution layers each is followed by an activation, and pooling layer. Also, in our network single dense layer with 512 neurons is used instead of multiple dense layers with large number of neurons. This helped in reducing the system complexity in terms of system parameters. The main contributions of this study are given as:

i) A Multilayer-Spatial Convolutional Neural Network with low complexity (few parameters) is proposed that is able to accurately detect the Covid-19 disease, achieving significant detection accuracy and AUC.

ii) The previous studies are based on either X-ray images or CT scans for Covid-19 detection. But, we have used both chest X-ray images and CT scans in this study to effectively train the proposed network for Covid-19 detection.

iii) We have developed two diverse databases for X-ray images and CT scans. The first database contains 723 chest X-ray images whereas the second database contains 3228 chest CT scans. Both databases are freely available for further studies.

The remaining paper is organized as follows. The literature review is given in Section II. The proposed deep learning method for Covid-19 detection is discussed in Section III. Materials and methods are presented in Section IV. Results and discussions are presented in Section V. Finally, the conclusions are presented in Section VI.

## II. Literature Review

From the public health viewpoint, quick isolation of patients is vital for controlling this contagious disease [1]-[3] and the best possible use of on hand resources that rapidly befall insufficient and plagued by an exponentially increasing number of patients and protracted times of the treatment. Researchers and scientists of the different disciplines are working along with public health officials to comprehend Covid-19 pathogenesis. Jointly they are working with the policymakers to urgently develop strategies, vaccines and curing drugs to treat the deadly novel disease. Thoracic radiology evaluation is used to diagnose suspected patients of Covid-19 [21]. But, scientific methods to identify the virus inside human bodies through Machine Learning and Deep Learning using chest X-ray images and Computed Tomography (CT) scans are potential methods. Timely detection and diagnosis of the disease is important in the efforts to guarantee timely treatment. Recent studies demonstrated imaging patterns on the chest X-Ray images and CT scans of the patients diagnosed with Covid-19 as potential analysis tool. The analysis revealed bilateral lung opacities on 98% chest X-ray images and CT scans in the infected people in Wuhan city and uttered lobular and subsegmental regions of consolidation as the most usual findings [22].

Other studies demonstrated high rates of ground-glass opacities and consolidation, with a rounded morphology and peripheral lung distribution [23]. Recently, many conventional image processing and machine/deep learning methods are used to diagnose the diseases by classifying the digitized chest X-ray images [24]-[25]. Class decomposition of the Covid-19 as Covid and non-Covid with X-ray images is considered as one of the significant methods for diagnosing this contagious disease [26]-[28]. Quick detection of the Covid-19 can help controlling the transmission of disease and to monitor the chain of infections. Chest CT scans are more helpful to diagnose Covid-19 as compared to the Reverse-Transcription Polymerase Chain Reaction (RT-PCR) which is collected from the swab samples of the patients and showed 97.3% accuracy to classify Covid-19 [29]. Convolution Neural Networks (CNNs) are the most accepted methods which have revealed a great ability and high precision to construe Covid-19 classification with medical imaging (X-ray images or CT scans). A Covid-19 classification method for the pathogen-confirmed Covid-19 is proposed [30] by using CNNs which are based on the Inception Net. The network achieved 82.9% classification accuracy by using 453 CT scans of pathogen-confirmed Covid-19. A multi-class classification method is proposed [31] to detect Covid-19 by using a pre-trained ResNet-50 (DRE-Net). For the classification, 86 CT scans of non- Covid-19, 100 CT scans of bacterial pneumonia and 88 CT scans of Covid-19 are used and showed 86% classification accuracy for Covid-19. Chest X-ray images are used to detect the Covid-19 in [32]. In the proposed method, deep features have been extracted using CNN which are based on pre-trained ImageNET. In the last layer Support Vector Machine, SVMs, are used for classification. A multi-class classification method is proposed [27] using deep CNN, called COVID-Net. Chest radiography images are used to classify Covid-19 and non- Covid-19.

Several other studies have been carried out to highlight recent contributions to Covid-19 detection [33]-[36]. In [37] a deep CNN, called DeTraC is adapted and validated for Covid-19 chest X-ray images classification. The proposed method traced irregularities in the chest X-ray images and examined class boundaries by using class decomposition method. The proposed method showed 95.12% classification accuracy (97.91% sensitivity and 91.87% specificity) for Covid-19. A deep learning-based classification method is proposed in [38] to extract deep features applying ResNet152 to classify chest x-ray images of Pneumonia and Covid-19 patients. SMOTE has been applied to balance the imbalance data points of normal and Covid-19. The proposed method showed 97.31% classification accuracy on Random Forest and 97.7% using XGBoost predictive classifiers. Various models including Alexnet, Googlenet, and Restnet18 have been analyzed to detect the Covid-19 in [33]. A novel method for detecting Covid-19 is proposed using chest X-ray images. A binary classification is used to detect the Covid-19 and non- Covid-19 whereas multi-class classification is used to detect Covid-19, non-Covid-19 and Pneumonia. The DarkNet was applied as a classifier for You Only Look Once (YOLO) real-time object detection system with 17 Convolutional layers using different filtering on each layer. The method showed 98.08% classification accuracy for binary classes and 87.02% for multi-class. An intelligent computer vision method called Residual Exemplar Local Binary Pattern (ResExLBP) has been proposed in [39] to detect Covid-19 which is based on preprocessing, feature extraction and feature selection, respectively. During the preprocessing, image-resizing and grayscale-conversion has been used whereas an

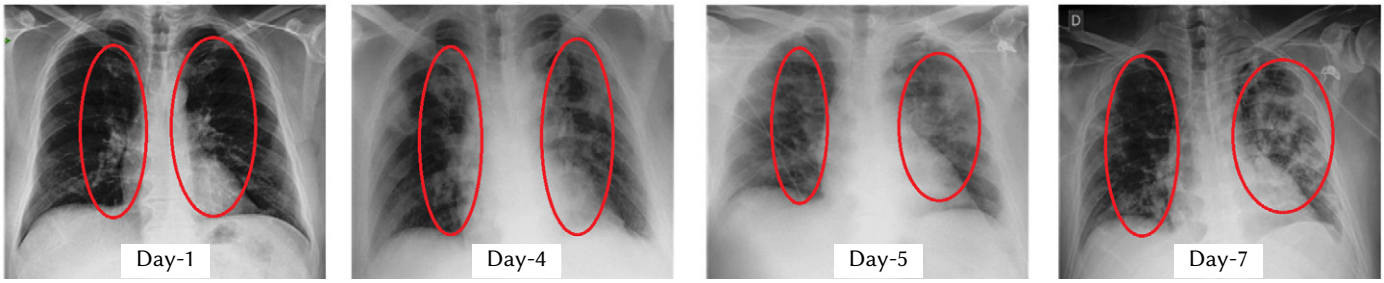



Fig. 1. Chest X-ray images of a 50-year-old COVID-19 patient over a week.

TABLE I. Comparative Analysis of Various Methods for Covid-19 Detection with Datasets and Evaluating Metrics

| S. No | Reference | Database Nature | Model Evolution Metrics | Network for Detection |
|---|---|---|---|---|
| 1 | [26] | **X-Ray Images**: Covid-19, Normal, Viral Pneumonia and Bacterial Pneumonia | AUC, Precision, NPV, F1-Score and Sensitivity | ResNet-50 with 50 Layers |
| 2 | [27] | **X-Ray Images**: Covid-19, Normal, and Viral Pneumonia | AUC, Precision, and Sensitivity | COVID-Net CNN |
| 3 | [30] | **CT Scans**: Covid-19 and Normal | AUC, Precision, NPV, F1-Score and Youden Index. | Fully connected CNN with Multiple Classifiers. |
| 4 | [31] | **CT Scans**: Covid-19, Normal, and Bacterial Pneumonia | AUC, and Recall (Sensitivity) | Details Relation Extraction neural network (DRE-Net) |
| 5 | [32] | **X-Ray Images**: Covid-19, Normal, and Viral Pneumonia | Accuracy, Sensitivity and Specificity | Deep CNN Architecture |
| 6 | [33] | **X-Ray Images**: Covid-19, Normal, Viral Pneumonia and Bacterial Pneumonia | Accuracy, Specificity, Recall, F1-score and Precision | Deep Transfer Learning CNN |
| 7 | [34] | **X-Ray Images**: Covid-19, and Viral Pneumonia | Accuracy, Specificity, Sensibility | Single Shot Multibox Detector (SSD) |
| 8 | [35] | **CT Scans**: Covid-19 | Diagnosis based detection | CNN and Management of Patients |
| 9 | [36] | **CT Scans**: Covid-19, Normal, and Viral Pneumonia | Accuracy, Sensitivity and Specificity | Multiple CNN with Classifiers |
| 10 | [37] | **X-Ray Images**: Covid-19, Normal, and SARS | Accuracy, Sensitivity and Specificity | Deep Transfer Learning CNN |

iterative ReliefF (IRF)-based feature selection is used. Decision Tree, Linear Discriminant (LD), Support Vector Machine (SVM), K-Nearest Neighborhood (KNN) and Subspace Discriminant (SD) approaches have been selected as classifiers during the classification phase. Zhao et al. [40] not only found ground-glass opacities (GGO) or mixed GGO in most of the patients, but they also observed a consolidation, and vascular dilation in the lesion. Li and Xia [35] reported GGO and consolidation, interlobular septal thickening and air bronchogram sign, with or without vascular expansion, as common CT features of Covid-19 patients. Peripheral focal or multifocal GGO affecting both lungs in 50%– 75% of patients are another observation [41]. Similarly, Zu et al. [42] and Chung et al. [43] discovered that 33% of chest CT scans can have rounded lung opacities. Fig. 1 shows chest X-ray images at days 1, 4, 5 and 7 for a 50-year-old Covid-19 patient.

In this paper, a deep learning model is proposed which is based on the 2D-Spatial Convolutional Neural Network for automatic detection of Covid-19 using chest X-ray images and CT scans. The proposed model is trained with 723 x-ray images and 3228 CT scans of both genders and various age groups. The x-ray images and CT scans are associated to Covid-19 and non-Covid-19 diagnosed patients. The proposed model provided an improved automated accurate detection of Covid-19 disease. Table I presents various deep learning methods with network types, database type and evaluation metrics used to assess the detection capabilities. It is clear from the Table I that most of the networks are complex and operate with more variables which make them complex as compared to the proposed method which has a relatively small number of parameters. Moreover, none of them has used both x-ray images and scans for detection. On the other hand, the proposed model has used x-ray images and scans for Covid-19 detection with less complexity.

## III. Proposed Deep Learning Method for Covid-19 Detection

CNN is as an effective machine learning method which provides up to date results by considering various layers of features. Recently 2D-CNN gained popularity in the area of image characterization [44], object detection and localization [45]-[47], face recognition [48], activity recognition [49]-[50]. Inspired by the performance of the 2D-CNN in the area of computer vision, we have used this network for automated detection of novel corona virus. In this study, a multilayer spatial CNN (2D-CNN) has been introduced to learn the prominent features needed for effective detection of Novel Covid-19 from X-ray images/CT Scan. CNN is a multilayer network architecture inspired from the neurobiology of the visual cortex. It contains an input layer, hidden layers, and an output layer. The hidden layer comprises of combination of the convolution layer, activation layers, pooling layers, normalization layers and fully connected layers. The

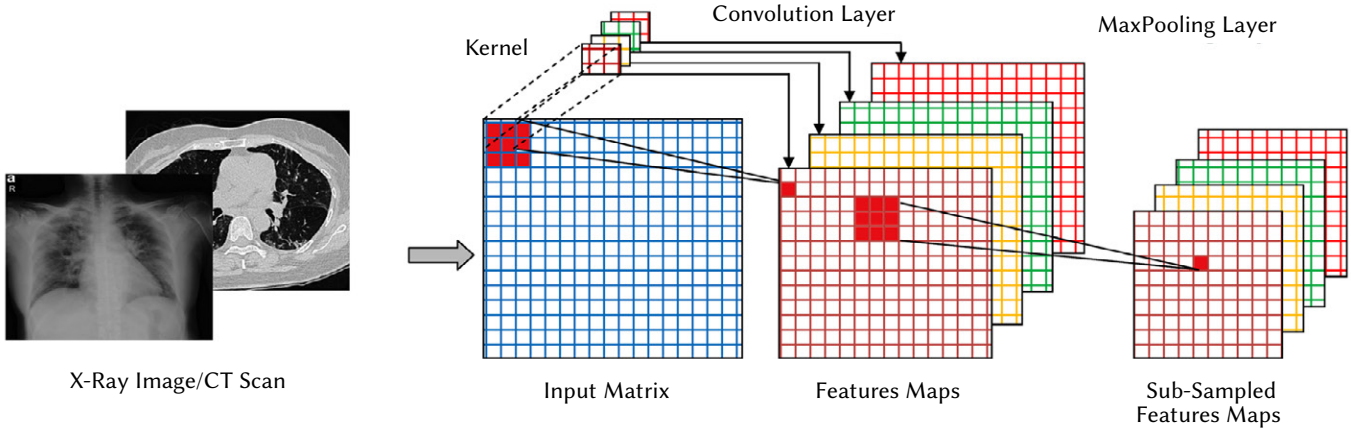



Fig. 2. Schematic presentation of Convolution and Max-Pooling layers.

convolution layers are used for extracting prominent features needed for classification of input data into desired classes. The convolution layer is the main building block of a CNN architecture. The prominent features are obtained through filters in the convolution layer. The filter coefficients convolved over height and width of the input data results in a 2D activation map of the filter. CNN has the capability to learn those filter coefficients, which activate when a particular feature at some spatial position is observed. The convolution layer is followed by an activation layer which is used to transform the input signal to an output signal. The output signal will be used as an input signal to the following layer. The activation layer normally uses a nonlinear function like sigmoid, tanh, ReLU, Leaky ReLU, etc. To speed up the learning process and avert the overfitting problem pooling layers are introduced in the CNN. The main task of this layer is to down sample the input data which reduces the spatial information to be processed. Among various pooling techniques average pooling and max pooling are the most prominent ones. The fully connected layer is similar to the conventional ANN. Its task is to set a path for the effective detection/classification.

A schematic presentation for the flow of input data from the convolution layer (C) and Max-pooling (M) layer, respectively, is given in Fig. 2. Inspired by the performance of CNN, a spatial CNN model has been proposed for auto mated detection of the COVID-19. The proposed model is composed of 5 Convolutional layers (with different number of filters, sizes, and strides), 5 maxpooling layers, a fully connected layer with 512 neurons, and a softmax classifier. An activation function has been used after each convolution layer and fully connected layer. For the activation function two different settings i.e. ReLU and Leaky ReLU activation functions are separately analyzed. The orientation of various layers used in the proposed model is depicted in Fig. 3. The first Convolutional layer contains 64 filters, each with size of (3, 3), and stride (1, 1). Similarly, the 2nd and 3rd Convolutional layer contain 128 filters each with size of (3, 3), and stride (1, 1). Furthermore, 4th and 5th Convolutional layer contain 256 filters each of size (3, 3), and stride (1, 1). All pooling layers use maxpooling strategy with the pooling window of size (2, 2), and strides (2, 2). The output of the last maxpooling layer is converted from 2D to 1D using a flatten layer. Then the output of the flatten layer is fed to the fully connected (Dense) layer with 512 neurons using sigmoid as an activation function. The fully connected layer is an actually conventional ANN architecture. At the output layer a softmax classifier is used to assign detection probabilities to each output. We have used SGD optimizer for learning weights. We have used a learning rate of 0.001, momentum = 0.9, and binary cross entropy loss function. The layer details and layer parameters of the model are given in Table II. First, the images are resized and preprocessed to fit in the

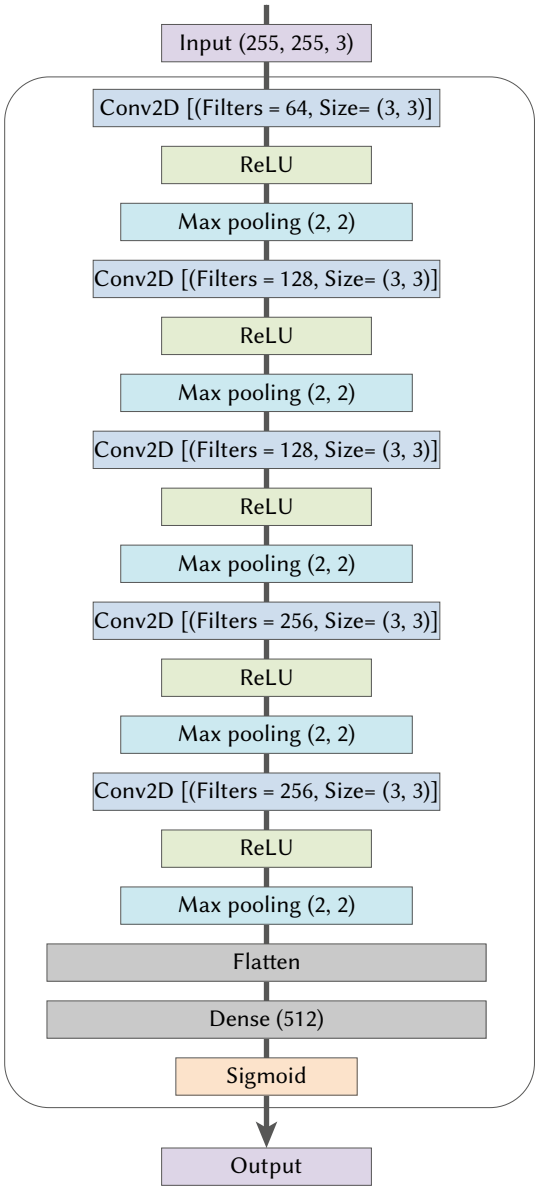



Fig. 3. The orientation of various layers used in the proposed model.

model. The features are extracted from the input images which are needed for the classification of the input data into desired classes. The features are obtained using filters in the convolution layer. The filter

coefficients convolved over height and width of the input data results in a 2D activation map of the filter. The convolution layer is followed by the activation layer which is used to transform the input signal to an output signal. The output signal is used as an input to next layer. The activation layer used nonlinear functions ReLU and Leaky ReLU. To boost the learning process and prevent the overfitting problem, the max pooling layers are used in the CNN. The task of this layer is to down sample the input data that minimizes the spatial information need to be processed.

TABLE II. The Layers and Parameters of the Proposed Model

| Layer Type | Output Shape | Parameters |
|---|---|---|
| Input Layer | [254 254 3] | 0 |
| Conv2D | [254 254 64] | 1792 |
| Maxpooling2 | [127 127 64] | 0 |
| Conv2D | [127 127 128] | 73856 |
| Maxpooling2 | [64 64 128] | 0 |
| Conv2D | [64 64 128] | 147584 |
| Maxpooling2 | [32 32 128] | 0 |
| Conv2D | [32 32 256] | 295168 |
| Maxpooling2 | [16 16 256] | 0 |
| Conv2D | [16 16 256] | 590080 |
| Maxpooling2 | [8 8 256] | 0 |
| Flatten | [16384] | 0 |
| Dense | [512] | 8389120 |
| Dropout | [512] | 0 |
| Dense | [2] | 1026 |
| Activation | [2] | 0 |

The main target of the proposed work is the efficient detection of Covid-19 in x ray images and CT scans. It is concluded from the literature that the performance of a relatively simple model such as VGG-16 [58] is better than the state of the art ResNet [26] and GoogleNet [57]. Therefore, this shows that the detection problem can be done with a relatively simple method. In this study, we focused on trying to efficiently detect the Covid-19 with a simple version of 2D-CNN inspired from VGG-11. Our network is a subset of VGG-11 which consists of 5-convolution layers, each is followed by ReLU/ Leaky ReLU activation function and max-pooling layer. Moreover, in our network, we have used a single dense layer that consists of 512 neurons instead of three dense layers with large number of neurons in each. Such network architecture arrangements helped us in reducing the system complexity in terms of system parameters and provided better results compared to the other networks. It is observed and verified that the proposed network outperformed the existing state of the art by considerable margins. So, our main and important contribution is to select a simple combination of different layers for achieving an efficient model in terms of system parameters and performance.

## IV. Experimental Setup

In the experimental setup, we discuss the database of x-ray images and scans to detect covid-19 in the patients. We have used several evaluation metrics to assess the effectiveness of the 2D-CNN-based learning method for covid-19 detection.

### *A. X-ray Images and CT Scans Databases*

To detect COVID-19 infection, we have used X-ray images from two different sources. For simplicity, we will use images instead of X-ray images afterward. The databases are developed by using images from various open access sources [51]-[53]. The first database contains a total of 625 images in which 125 images belong to Covid-19 diagnosed patients and 500 are normal images. Similarly, the second database contains a total of 98 images in which 70 images belong to Covid-19 diagnosed patients and 28 are normal images. In our study, we have combined both databases and generated a new diverse database with 723 images of both genders and various age groups in which 195 images belong to Covid-19 diagnosed patients and 528 are non-Covid-19 images. We also have used Computed Tomography (CT) scans from two different sources to detect Covid-19. For simplicity, we will use the term scans instead of CT scans afterward. The databases are developed using scans from the Tongji Hospital, Wuhan, China [54] and Sao Paulo, Brazil [55]. The first dataset from Tongji Hospital, Wuhan contains a total of 746 scans where 349 scans are associated to Covid-19 patients whereas 397 are non-Covid-19 scans. Similarly, the second database from Sao Paulo, Brazil contains a total of 2482 scans in which 1252 scans belong to Covid-19 patients and 1230 are non-Covid-19 scans. We have combined both databases and generated a new diverse database with a total of 3228 scans of both genders and various age groups. The new database contains 1601 scans which belong to Covid-19 patients whereas 1627 are non-Covid-19 scans. Fig. 4 demonstrates samples images and scans of Covid-19 and non-Covid-19 cases selected from the new databases.

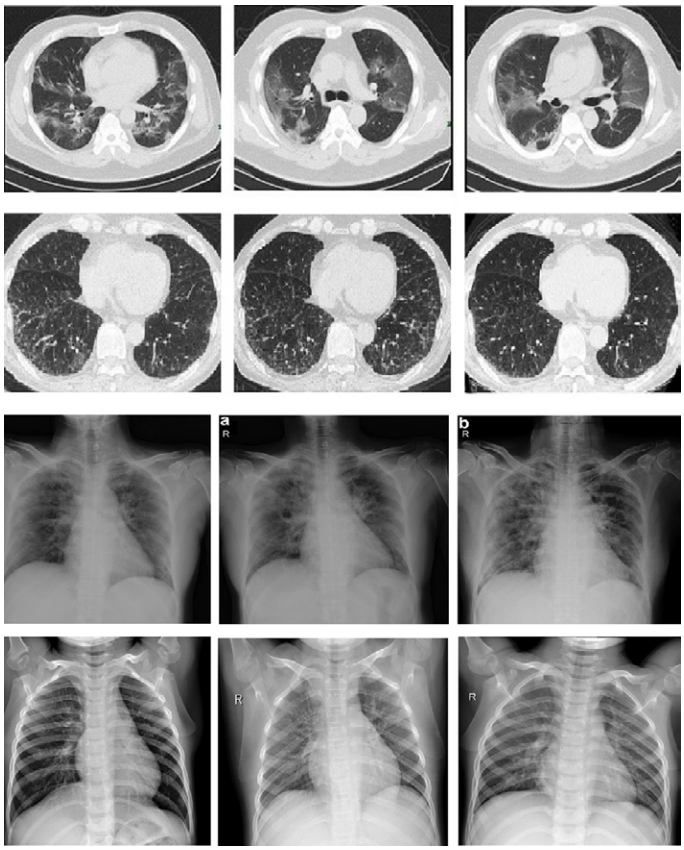


Fig. 4. Samples images and scans of COVID-19 and non-COVID-19 cases in the new databases.

### *B. Evaluation Criteria*

To examine the effectiveness of the proposed model, the confusion matrix along with the Receiver operating characteristics (ROC) and Area under Curve (AUC) [56] are calculated, which determines the potentials of the proposed model for Covid-19 detection. The usefulness and productivity of the proposed model are also measured using the conventional evaluation metrics including accuracy,

precision, sensitivity, and F1 score, which are represented in terms of the confusion matrix. The evaluation metrics are given by the following equations as:

$$\text{Accuracy}: \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Sensitivity}: \frac{TP}{TP+FN} \quad (2)$$

$$\text{Precision}: TP/(TP+FP) \quad (3)$$

$$\text{F1-Score}: \frac{2*TP}{2*TP+FP+FN} \quad (4)$$

Where, *TP*, *TN*, *FP*, and *FN* denote True Positive, True Negative, False Positive, and False Negative, respectively.

## V. Results and Discussions

We performed a number of intense experiments to detect Covid-19 using the two new diverse databases containing CT scans and X-ray images. We have trained the MS*Cov*CNN deep learning model to classify CT scans and X-ray images into Covid-19 and non-Covid-19 cases. The performance of the proposed model is examined using random validation procedure for the binary classification problem. We have performed experiments by splitting the training data in two splits: 80:20 and 50:50, that is, 80% of CT scans and X-ray images are used for training and 20% for validation. Similarly, 50% of CT scans and X-ray images are used for training and 50% for validation. In the experiments, we have used two activation functions: ReLU and Leaky ReLU and repeated the experiments for both split separately. Table III shows the experimental results in terms of the Accuracy and AUC for the two splits using ReLU and Leaky ReLU activations. It can be observed from Table III that a high average network accuracy and AUC for CT scans and X-ray images are achieved when leaky ReLU is used in the proposed model. The average accuracy of the network is improved by 1.16% and 1.06% for X-ray images and CT scans, respectively. Similarly, the average AUC of the network is improved by 2.01% for the X-ray images and 0.47% for CT scans. Consequently, the leaky ReLU is selected as potential activation function for the proposed model. At the start of training procedure, we observed a significant increase in the values of loss function which has largely been decreased at the end of the training procedure. When the proposed deep learning model examined all X-ray images and CT scans over and over again for all epochs during the training, the rapid ups and downs are slowly reduced in the later part of the training.

TABLE III. Accuracy and AUC of the Proposed Model for Chest X-ray Images and CT Scans Using ReLU and Leaky ReLU Activations

| Database: Chest X-Ray Images | | | | |
|---|---|---|---|---|
| Data Split | ReLU | | Leaky ReLU | |
| | Accuracy | AUC | Accuracy | AUC |
| 80:20 | 90.76% | 96.42% | 91.53% | 97.11% |
| 50:50 | 93.59% | 95.33% | 95.72% | 98.65% |
| Average | 92.18% | 95.87% | 93.34% | 97.88% |
| **Database: Chest CT Scans** | | | | |
| Data Split | ReLU | | Leaky ReLU | |
| | Accuracy | AUC | Accuracy | AUC |
| 80:20 | 91.95% | 95.79% | 92.64% | 96.31% |
| 50:50 | 88.82% | 94.52% | 90.25% | 95.52% |
| Average | 90.38% | 95.15% | 91.44% | 95.92% |

Tables IV-V indicate the performance of the proposed deep learning model for two splits using chest X-ray images and CT scans. The proposed model achieved significant results in terms of the Covid-19 detection and achieved improved accuracy percentage along with other important metrics. It can be observed from Tables IV-V that the proposed model achieved better results for chest X-ray images as compared to CT scans. The proposed model achieved 91.53% network accuracy for 80:20 split whereas achieved 95.72% network accuracy for 50:50 split. A high accuracy is reported for 50:50 split setting. The AUC, an important evaluation parameter indicates that the proposed model achieved better results. An average of 97.88% AUC is achieved with the proposed model. Moreover, 50:50 split achieved better AUC percentage as compared to the 80:20 split for chest X-ray images. Similarly, 91.44% average network accuracy and 95.92% AUC for CT scans are achieved with the proposed model. We secondly examined the results of the proposed model by using Confusion Matrixes and ROC for the binary classification problem in order to detect the novel Covid-19. The Confusion Matrixes and ROC are drawn for 80:20 and 50:50 splits of X-ray images and CT scans for both ReLU and Leaky ReLU activation functions. The vertical axis of confusion matrix shows the true labels whereas horizontal axis indicates the predicted labels of Covid-19 and non-Covid-19, respectively. For example, consider the confusion matrix obtained from the 80:20 split of X-ray images for ReLU activation function, see Fig. 5(A). The element in first-row first-column indicates true negatives which means that 98% of negative samples are classified correctly (non-Covid-19). Similarly, the element of first-row second-column indicates false positive which means that 2% of negative samples are confused with the positive labels. The element of second-row first-column represents false negative which means that 19% of positive labels are identified as negative labels. Finally, the element of second -row and second-column shows true

TABLE IV. Performance Evaluation of the Proposed Model: S*Cov*CNN Using Accuracy, Sensitivity and AUC

| Database: X-Ray Images | | | | | | | |
|---|---|---|---|---|---|---|---|
| Split | TP | TN | FP | FN | Accuracy | Sensitivity | AUC |
| 80:20 | 32 | 87 | 3 | 8 | 91.53% | 80% | 97.11% |
| 50:50 | 65 | 204 | 3 | 9 | 95.72% | 87.83% | 98.65% |
| Avg | 48.5 | 145.5 | 3 | 8.5 | 93.63% | 84% | 97.88% |
| **Database: CT Scans** | | | | | | | |
| Split | TP | TN | FP | FN | Accuracy | Sensitivity | AUC |
| 80:20 | 282 | 259 | 35 | 8 | 92.63% | 97.24% | 96.31% |
| 50:50 | 541 | 597 | 48 | 75 | 90.24% | 87.82% | 95.52% |
| Avg | 411.5 | 428 | 41.5 | 41.5 | 91.44% | 92.53% | 95.92% |

TABLE V. Performance Evaluation of the Proposed Model: S*Cov*CNN Using Precision and F1-score

| Database: X-Ray Images | | | | | | |
|---|---|---|---|---|---|---|
| Split | TP | TN | FP | FN | Precision | F1 Score |
| 80:20 | 32 | 87 | 3 | 8 | 91.42% | 85.33% |
| 50:50 | 65 | 204 | 3 | 9 | 95.58% | 91.54% |
| Avg. | 48.5 | 145.5 | 3 | 8.5 | 93.50% | 88.44% |
| **Database: CT Scans** | | | | | | |
| Split | TP | TN | FP | FN | Precision | F1 score |
| 80:20 | 282 | 259 | 35 | 8 | 88.95% | 92.91% |
| 50:50 | 541 | 597 | 48 | 75 | 91.85% | 89.79% |
| Avg. | 411.5 | 428 | 41.5 | 41.5 | 90.40% | 91.35% |

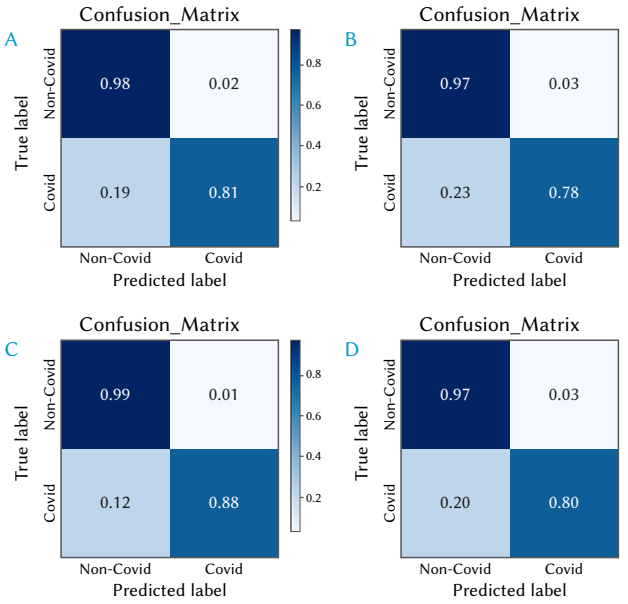



Fig. 5. Confusion Matrices for Chest X-ray images. (A) 80:20 split with ReLU activation function, (B) 80:20 split with Leaky ReLU activation function, (C) 50:50 split with ReLU activation function, (D) 50:50 split with Leaky ReLU activation function.

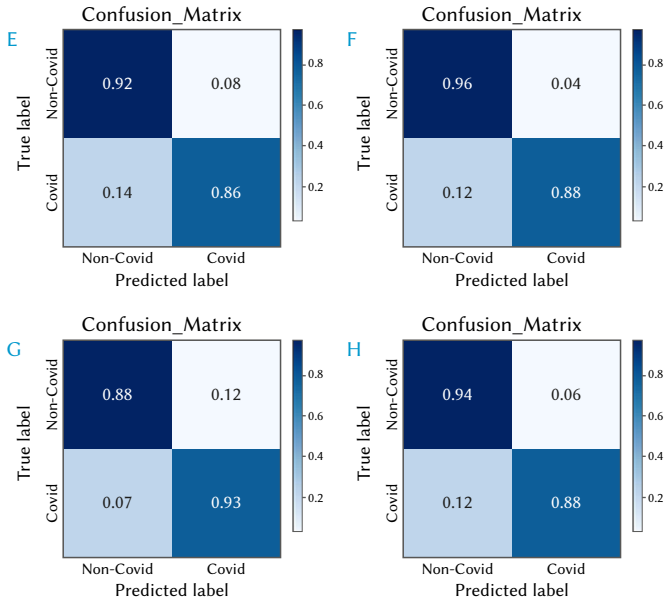



Fig. 6. Confusion Matrices for Chest CT scans. (E) 80:20 split with ReLU activation function, (F) 80:20 split with Leaky ReLU activation function, (G) 50:50 split with ReLU activation function, (H) 50:50 split with Leaky ReLU activation function.

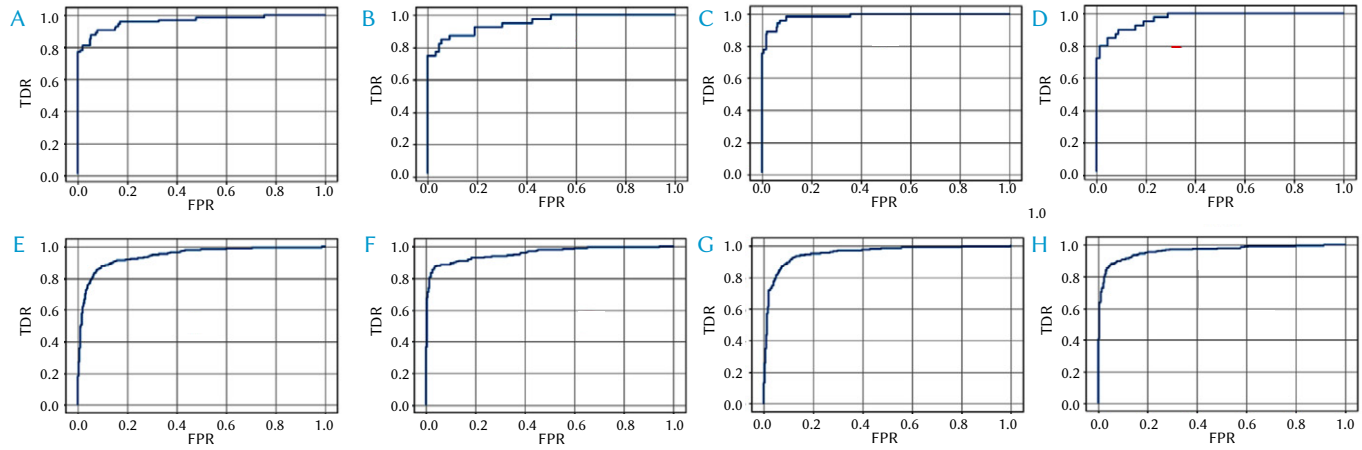



Fig. 7. ROC analysis for Chest X-ray images and CT scans. (A) 80:20 split with ReLU activation function, (B) 80:20 split with Leaky ReLU activation function, (C) 50:50 split with ReLU activation function, (D) 50:50 split with Leaky ReLU activation function, (E) 80:20 split with ReLU activation function, (F) 80:20 split with Leaky ReLU activation function, (G) 50:50 split with ReLU activation function, (H) 50:50 split with Leaky ReLU activation function.

positive which means that 81% of the positive samples are correctly classified as Covid-19. Consider the confusion matrix obtained from the 50:50 split of CT scans for leaky ReLU activation function, see Fig. 6(H). The element in first-row first-column indicates true negatives which means that 94% of negative samples are classified correctly (non-Covid-19). Similarly, the element of first-row second-column indicates false positive which means that 6% of negative samples are confused with the positive labels. The element of second-row first-column represents false negative which means that 12% of positive labels are identified as negative labels. Finally, the element of second-row and second-column shows true positive which means that 88% of positive samples are correctly classified as Covid-19. The confusion matrix for chest X-ray images and CT scans are illustrated in Fig. 5-6, respectively. Non-linear filters in the initial layers of network act as preprocessing layers which helps in extracting prominent features by learning the filter coefficient. So, preprocessing in case of convolution neural network may not help in improving the results. Comparison of the complexity of state of the art networks is given in Table VI. ROC plots are depicted in Fig. 7 which indicates the true positive vs. false positive rates. ROC plots are used to show the separation of features from each other. We also provided log loss, MSE, MAE, and MLSE for evaluating the proposed method in Table VII.

TABLE VI. Network Complexity Analysis

| S.No. | Technique | Parameters |
|---|---|---|
| 1. | AlexNet | 62 Million |
| 2. | VGG 16 | 138.36 Million |
| 3. | Inception V3 | 41.33 Million |
| 4. | ResNet 50 | 25.56 Million |
| 5. | Proposed Method | 9.49 Million |

TABLE VII. Comparison of Various Methods for Loss

| Method | Split | MSE | Log Loss | MAE | MSLE |
|---|---|---|---|---|---|
| Leaky ReLU | 50:50 | 0.1517 | 0.2719 | 0.1809 | 0.1235 |
| Leaky ReLU | 80:20 | 0.1369 | 0.1995 | 0.1577 | 0.1215 |
| ReLU | 50:50 | 0.1819 | 0.3404 | 0.2099 | 0.1544 |
| ReLU | 80:20 | 0.1396 | 0.2100 | 0.1606 | 0.12298 |

### A. Comparison with Other Methods

In this section, we have compared the proposed deep learning model with other competing deep learning models for Covid-19 detection. For comparison purpose, we have selected xDNN [55], ResNet [26], GoogleNet [57], VGG-16 [58], AlexNet [57], Decision Tree [59], and AdaBoost [60]. All deep learning approaches for Covid-19 detection are evaluated using Accuracy, precision, sensitivity, F1-score and AUC. Table VIII shows the performance of the proposed deep learning model and the competing models. In this experiment we have combined X-rays and CT scans into a single dataset. We achieved better performance in terms of Accuracy, precision, sensitivity, F1-score and AUC compared to other competing methods for covid-19 detection in the literature. For example, accuracy of the proposed detection method is improved from 91.73%, 93.75 and 94.96% with GoogleNet, AlexNet and ResNet to 97.48% with S*Cov*CNN. Similarly, the AUC is improved from 95.19%, 79.51%, 94.96% and 97.36% with the AdaBoost, Decision Tree, VGG-16 and xDNN to 97.36% with S*Cov*CNN. Precision, sensitivity and F1 score of the proposed model is consistently higher than the competing methods. Decision Tree performed less as compared to other methods. The improvements in the evaluation metrics with respect to Decision Tree is plotted in Fig. 8. In convolutional neural networks complexity of a model is defined by the number of parameters.

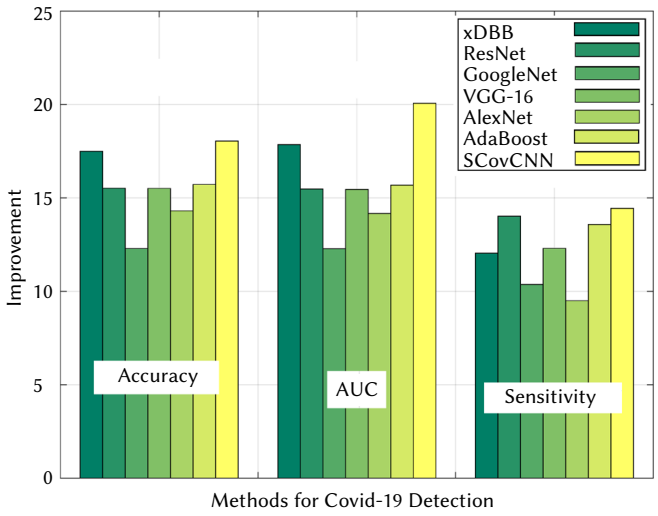



Fig. 8. Accuracy, AUC and Sensitivity improvements of various methods with reference to Decision Tree.

## VI. Conclusions

In this study, we have proposed a multilayer-Spatial Convolutional Neural Network for automatic detection of Covid-19 using chest X-ray images and CT scans. The proposed model showed 98.18% detection accuracy and 99.98% AUC for chest x-ray images and 97.14% detection accuracy and 99.51% AUC for chest CT scans. The previous studies are based on either X-ray images or CT scans for Covid-19 detection. But, we have used both chest X-ray images and CT scans in this study to effectively train the proposed network for Covid-19 detection. We have developed two diverse databases for X-ray images and CT scans. First database contains 723 chest X-ray images whereas the second database contains 3228 chest CT scans. Both databases are freely available for further studies. The proposed model is evaluated using a number of metrics including confusion matrix, ROC, AUC, accuracy, precision, sensitivity, and F1 scores, respectively. We have performed experiments by splitting the training data in two splits: 80:20 and 50:50, that is, 80% of CT scans and X-ray images are used for training and 20% for validation. Similarly, 50% of CT scans and X-ray images are used for training and 50% for validation. We have drawn the following conclusions:

1. The average accuracy and AUC of the proposed model is improved by 1.16% and 1.06% for X-ray images and CT scans whereas 2.01% for X-ray images and 0.47% for CT scans. Therefore, it is concluded that the leaky ReLU is the potential activation function for the proposed model.
2. We concluded that there was a significant increase in the values of loss function which has largely been decreased at the end of training procedure. The proposed deep learning model examined all X-ray images and CT scans over and over again for all epochs during the training, hence, rapid fluctuations in loss function values are slowly reduced in the later part of the training.
3. It is concluded that the proposed model achieved significant results in terms of the Covid-19 detection and achieved higher accuracy, AUC, sensitivity and F1 scores. The proposed model achieved 91.53% network accuracy for 80:20 split whereas achieved 95.72% network accuracy for 50:50 split. A high accuracy is reported for 50:50 split setting.
4. It is concluded that the proposed model achieved better performance in terms of the accuracy, precision, sensitivity, F1-score and AUC compared to competing methods for covid-19 detection. The accuracy of the proposed detection method is improved from 91.73%, 93.75 and 94.96% with GoogleNet, AlexNet and ResNet to 97.48% with S*Cov*CNN.

In the future work, we will be devoted in attempting further improvements in the performance of the proposed model and will extend the proposed model into a more powerful model. In addition, we will systematically examine the complex networks and classifiers to find more accurate results in terms of Covid-19 detection.

TABLE VIII Comparison with Competing Methods

| Database: X-Ray Images/CT Scans | | | | | |
|---|---|---|---|---|---|
| Methods | Accuracy | Precision | Sensitivity | F1 score | AUC |
| xDNN [55] | 97.38% | 91.6% | 95.53% | 97.31% | 97.36% |
| ResNet [26] | 94.96% | 93.00% | 97.15% | 95.03% | 94.98% |
| GoogleNet [57] | 91.73% | 90.20% | 93.50% | 91.82% | 91.79% |
| VGG-16 [58] | 94.96% | 94.02% | 95.43% | 94.97% | 94.96% |
| AlexNet [57] | 93.75% | 94.98% | 92.28% | 93.61% | 93.68% |
| Decision Tree [59] | 79.44% | 76.81% | 83.13% | 79.84% | 79.51% |
| AdaBoost [60] | 95.16% | 93.63% | 96.71% | 95.14% | 95.19% |
| S*Cov*CNN | **97.48%** | **97.18%** | **97.57%** | **97.37%** | **99.57%** |

## References


[1] H. Lau, V. Khosrawipour, P. Kocbach, A. Mikolajczyk, H. Ichii, J. Schubert, T. Khosrawipour, "Internationally lost COVID-19 cases," *Journal of Microbiology, Immunology and Infection*, vol. 53, no. 3, 2020.

[2] J. F. Zhang, K. Yan, H. H. Ye, J. Lin, J. J. Zheng, T. Cai, "SARS-CoV-2 turned positive in a discharged patient with COVID-19 arouses concern regarding the present standard for discharge," *International Journal of Infectious Diseases*, vol. 97, pp. 212-214, 2020.

[3] M. Dur-e-Ahmad, M. Imran, "Transmission Dynamics Model of Coronavirus COVID-19 for the Outbreak in Most Affected Countries of the World," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 2, pp. 7-10, 2020.

[4] T. Singhal, "A review of coronavirus disease-2019 (COVID-19)," *The Indian Journal of Pediatrics*, vol. 87, no. 4, pp. 1-6, 2020.

[5] C. C. Lai, T.P. Shih, W.C. Ko, H.J. Tang, P.R. Hsueh, "Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and corona virus disease-2019 (COVID-19): the epidemic and the challenges," *International journal of antimicrobial agents,* vol. 55, no. 3, 105924, 2020.

[6] J. Li, J.J. Li, X. Xie, X. Cai, J. Huang, X. Tian, H. Zhu, Game consumption and the 2019 novel coronavirus, *The Lancet Infectious Diseases*, vol. 20, no. 3, pp. 275-276, 2020.

[7] J. M. Sharfstein, S.J. Becker, M.M. Mello, "Diagnostic testing for the novel coronavirus," *Jama*, vol. 323, no. 15, pp. 1437-1438, 2020.

[8] L. Chang, Y. Yan, L. Wang, "Coronavirus disease 2019: coronaviruses and blood safety," *Transfusion medicine reviews*, vol. 34, no. 2, pp. 85-80, 2020.

[9] M. A. Shereen, S. Khan, A. Kazmi, N. Bashir, R. Siddique. "COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses," *Journal of Advanced Research*, vol. 24, pp. 91-98, 2020.

[10] F. A. Rabi, M. S. Al Zoubi, G.A. Kasasbeh, D.M. Salameh, A.D. Al-Nasser, "SARS-CoV-2 and coronavirus disease 2019: what we know so far," *Pathogens*, vol. 9, no. 3, 231, 2020.

[11] A. York, "Novel coronavirus takes flight from bats?," *Nature Reviews Microbiology*, vol. 18, no. 4, pp. 191-191, 2020.

[12] T. T. Y. Lam, M. H. H. Shum, H. C. Zhu, Y. G. Tong, X. B. Ni, Y.S. Liao,... G.M. Leung, "Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins," *Nature*, vol. 583, no. 7815, pp. 282-285, 2020.

[13] M. Giovanetti, D. Benvenuto, S. Angeletti, M. Ciccozzi, "The first two cases of 2019-nCoV in Italy: Where they come from?," *Journal of medical virology*, vol. 92, no. 5, pp. 518-521, 2020.

[14] M. L. Holshue, C. DeBolt, S. Lindquist, K. H. Lofy, J. Wiesman, H. Bruce, G. Diaz, "First case of 2019 novel coronavirus in the United States," *New England Journal of Medicine,* vol. 382, pp. 929-936, 2020.

[15] A. Bastola, R. Sah, A.J. Rodriguez-Morales, B.K. Lal, R. Jha, H.C. Ojha,... K. Morita, "The first 2019 novel coronavirus case in Nepal," *The Lancet Infectious Diseases*, vol. 20, no. 3, pp. 279-280, 2020.

[16] C. Rothe, M. Schunk, P. Sothmann, G. Bretzel, G. Froeschl, C. Wallrauch, M. Seilmaier, "Transmission of 2019-nCoV infection from an asymptomatic contact in Germany," New England Journal of Medicine, vol. 382, no. 10, pp. 970-971, 2020.

[17] L. T. Phan, T. V. Nguyen, Q.C. Luong, T.V. Nguyen, H.T. Nguyen, H.Q. Le, Q.D. Pham, "Importation and human-to-human transmission of a novel coronavirus in Vietnam," *New England Journal of Medicine,* vol. 382, no. 9, pp. 872-874, 2020.

[18] Coronavirus (COVID-19) Map. Available online: https://www.who.int/ (accessed on 6 June 2020).

[19] World Health Organization; Coronavirus disease (COVID-19) advice for the public. Accessed on: April 11, 2020, https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public

[20] Centers for Disease Control and Prevention; 2019 Novel Coronavirus (2019-nCoV). Accessed on: April 11, 2020, https://www.cdc.gov/coronavirus/2019-ncov/index.html

[21] H. Shi, X. Han, N. Jiang, Y. Cao, O. Alwalid, J. Gu ... C. Zheng, "Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study," *The Lancet Infectious Diseases,* vol. 20, no. 4, pp. 424-434, 2020.

[22] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, ... Z. Cheng, "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," *The lancet*, vol. 395, no. 10223, pp. 497-506, 2020.

[23] T. Lupia, S. Scabini, S. M. Pinna, G. Di Perri, F.G. De Rosa, S. Corcione, "2019-novel coronavirus outbreak: A new challenge," *Journal of Global Antimicrobial Resistance*, vol. 21, pp. 22-27, 2020.

[24] T, Rahmat, A. Ismail, S. Aliman, "Chest X-Rays Image Classification in Medical Image Analysis," *Applied Medical Informatics,* vol. 40, no. 3-4, pp. 63-73, 2018.

[25] I. M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, A. Saalbach, "Comparison of deep learning approaches for multi-label chest X-ray classification," *Scientific reports,* vol. 9, no.1, pp. 1-10, 2019.

[26] M. Farooq, A. Hafeez, "Covid-resnet: A deep learning framework for screening of covid19 from radiographs," arXiv preprint arXiv:2003.14395, 2020.

[27] L. Wang, A. Wong, "COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images," arXiv, arXiv-2003, 2020.

[28] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong ... K. Cao, "Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT," *Radiology*, vol. 296, no. 2, 200905, 2020.

[29] T. Ai, Z. Yang, H. Hou, C. Zhan, C. Chen, C., W. Lv, ... L. Xia, "Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases," *Radiology,* vol. 296, no. 2, pp. E32-E40, 2020.

[30] S. Wang, B. Kang, J. Ma, X. Zeng, M. Xiao, J. Guo, ... B. Xu, "A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19)," MedRxiv.2020.

[31] Y. Song, S. Zheng, L. Li, X. Zhang, X. Zhang, Z. Huang, ... Y. Chong, "Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images," medRxiv, 2020.

[32] P.K. Sethy, S.K. Behera, "Detection of coronavirus disease (covid-19) based on deep features," Preprints, 2020030300, 2020.

[33] A. Narin, C. Kaya, Z. Pamuk, "Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks," arXiv preprint arXiv:2003.10849, 2020.

[34] F.A. Saiz, I. Barandiaran, "COVID-19 Detection in Chest X-ray Images using a Deep Learning Approach," *International Journal of Interactive Multimedia and Artificial Intelligence,* vol. 6, no. 2, pp. 11-14, 2020.

[35] Y. Li, L. Xia, "Coronavirus disease 2019 (COVID-19): role of chest CT in diagnosis and management," *American Journal of Roentgenology*, vol. 214, no. 6, pp. 1280-1286, 2020.

[36] X. Xu, X. Jiang, C. Ma, P. Du, X. Li, S. Lv, L. Yu, ... L. Li, "A Deep Learning System to Screen Novel Coronavirus Disease 2019 Pneumonia," *Engineering,* vol. 6, no. 10, pp. 1122-1129, 2020.

[37] A. Abbas, M.M. Abdelsamea, M. M. Gaber, "Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network," arXiv preprint arXiv:2003.13815, 2020.

[38] R. Kumar, R. Arora, V. Bansal, V.J. Sahayasheela, H. Buckchash, J. Imran, ... B. Raman, "Accurate Prediction of COVID-19 using Chest X-Ray Images through Deep Feature Learning model with SMOTE and Machine Learning Classifiers," medRxiv. 2020.

[39] T. Tuncer, S. Dogan, F. Ozyurt, "An automated Residual Exemplar Local Binary Pattern and iterative ReliefF based corona detection method using lung X-ray image," *Chemometrics and Intelligent Laboratory Systems*, vol. 203, 104054, 2020.

[40] W. Zhao, Z. Zhong, X. Xie, Q. Yu, J. Liu, "Relation between chest CT findings and clinical conditions of coronavirus disease (COVID-19) pneumonia: a multicenter study," *American Journal of Roentgenology*, vol. 214, no. 5, pp. 1072-1077, 2020.

[41] J.P. Kanne, B.P. Little, J. H. Chung, B.M. Elicker, L.H. Ketai, "Essentials for radiologists on COVID-19: an update—radiology scientific expert panel," *Radiology*, vol. 296, no. 2, 2020.

[42] W. Kong, P.P. Agarwal, "Chest imaging appearance of COVID-19 infection," *Radiology: Cardiothoracic Imaging*, vol. 2, no. 1, e200028, 2020.

[43] Z. Y. Zu, M. D. Jiang, P.P. Xu, W. Chen, Q.Q. Ni, G.M. Lu, L.J. Zhang, "Coronavirus disease 2019 (COVID-19): a perspective from China," *Radiology*, vol. 296, no. 2, pp. E15-E25, 2020.

[44] P. N. Druzhkov, V.D. Kustikova, "A survey of deep learning methods and software tools for image classification and object detection," *Pattern Recognition and Image Analysis*, vol. 26, no. 1, pp. 9-15, 2016.

[45] X. Zhou, W. Gong, W. Fu, F. Du, "Application of deep learning in object detection," in *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS),* IEEE, May 2017, pp. 631-634.

[46] K. Nguyen, C. Fookes, A. Ross, S. Sridharan, "Iris recognition with off-the-shelf CNN features: A deep learning perspective," *IEEE Access*, vol. 6, no. 18848-18855, 2017

[47] S. Milyaev, I. Laptev, "Towards reliable object detection in noisy images,"

*Pattern Recognition and Image Analysis*, vol. 27, no. 4, pp. 713-722, 2017.

[48] R. Ranjan, S. Sankaranarayanan, A. Bansal, N. Bodla, J.C. Chen, V.M. Patel, ... R. Chellappa, "Deep learning for understanding faces: Machines may be just as good, or better, than humans," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 66-83, 2018.

[49] M. Papakostas, T. Giannakopoulos, F. Makedon, V. Karkaletsis, "Short-term recognition of human activities using convolutional neural networks," in *2016 12th international conference on signal-image technology & internet-based systems (SITIS)*, IEEE, January 2016, pp. 302-307.

[50] N. Yudistira, T. Kurita, "Gated spatio and temporal convolutional neural network for activity recognition: towards gated multimodal deep learning," *EURASIP Journal on Image and Video Processing*, vol. 2017, 85, 2017.

[51] J.P. Cohen, P. Morrison, L. Dao, K. Roth, T.Q. Duong, M. Ghassemi, "COVID-19 Image Data Collection: Prospective Predictions Are the Future," arXiv:2006.11988, https://github.com/ieee8023/COVID-chestxray-dataset/

[52] https://github.com/ieee8023/covid-chestxray-dataset

[53] T. Rahman, M. Chowdhury, A. Khandakar, "COVID-19 Radiography Database," https://www.kaggle.com/tawsifurrahman/covid19-radiography-database

[54] J. Zhao, Y. Zhang, X. He, P. Xie, "COVID-CT-Dataset: a CT scan dataset about COVID-19," arXiv preprint arXiv:2003.13865, 2020, https://github.com/UCSD-AI4H/COVID-CT

[55] E. Soares, P. Angelov, S. Biaso, M.H. Froes, D.K. Abe, "SARS-CoV-2 CT-scan dataset: A large dataset of real patients CT scans for SARS-CoV-2 identification," medRxiv, 2020.

[56] Y. Heryadi, H.L.H.S. Warnars, "Learning temporal representation of transaction amount for fraudulent transaction recognition using CNN, Stacked LSTM, and CNN-LSTM," in *2017 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, Phuket, 2017, pp. 84-89.

[57] C. Alippi, S. Disabato, M. Roveri, "Moving convolutional neural networks to embedded systems: the alexnet and VGG-16 case," in *2018 17th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, 2018, pp. 212-223.

[58] P. Ballester, R.M. Araujo, "On the performance of GoogLeNet and AlexNet applied to sketches," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016, pp. 1124-1128.

[59] M. A. Friedl, C. E. Brodley, "Decision tree classification of land cover from remotely sensed data," *Remote sensing of environment*, vol. 61, no. 3, pp. 399-409, 1997.

[60] T. Hastie, S. Rosset, J. Zhu, H. Zou, "Multi-class adaboost. *Statistics and its Interface*," vol. 2, no. 3, pp. 349-360, 2009.

[61] S. J. Fong, G. Li, N. Dey, R.G. Crespo, E. Herrera-Viedma, "Finding an accurate early forecasting model from small dataset: A case of 2019-ncov novel coronavirus outbreak," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 1, pp. 132-140, 2020.

[62] N. M. Nawi, W.H. Atomi, M.Z. Rehman, "The effect of data pre-processing on optimized training of artificial neural networks," vol. 11, pp. 32-39, 2013.

M. Irfan Khattak

Dr. M. Irfan Khattak is working as an Associate professor in the Department of Electrical Engineering in University of Engineering and Technology Peshawar. He did his B.Sc Electrical Engineering from the same University in 2004 and did his PhD from Loughborough University UK in 2010. After doing his PhD he was appointed as Chairman Electrical Engineering Department at UET Bannu Campus for five years and took care of the academic and research activities at the department. Later in 2016 he was appointed as Campus Coordinator of UET Kohat Campus and took the administrative control of the Campus. He is also heading a research group "Microwave and Antenna Research Group" where he is supervising Post grad Students working on Latest trends in Antenna Technology like 5G and Graphene Nano-antennas for Terahertz, Optoelectronic and Plasmonic Applications etc. His research interest involves Antenna Design, On-Body Communications, Anechoic Chamber Characterization, Machine Learning/Deep Learning and its applications including image processing and Speech processing and Speech Enhancement. Besides his research activities he is certified OBE Expert with Pakistan Engineering Council for organizing OBA based accreditation visits.

Mu'ath Al-Hasan

Mu'ath Al-Hasan received his B.A.Sc. degree in electrical engineering from the Jordan University of Science and Technology, Jordan, in 2005, the M.A.Sc in wireless communications from Yarmouk University, Jordan in 2008, and the Ph.D. degree in Telecommunication engineering from Institut National de la Recherche Scientifique (INRS), Université du Québec, Canada, 2015. From 2013 to 2014, he was with Planets Inc., California, USA. In May 2015, he joined Concordia University, Canada as postdoctoral fellowship. He is currently an Assistant Professor with Al Ain University, United Arab Emirates. His current research interests include antenna design at millimeter-wave and Terahertz, channel measurements in Multiple-Input and Multiple-Output (MIMO) systems, and Machine Learning and Artificial Intelligence in antenna design, Machine Learning/Deep Learning and its applications including image processing and Speech processing and Speech Enhancement, Antennas designing.

Atif Jan

Atif Jan obtained his B.Sc. degree in Electrical Engineering from University of Engineering and Technology (UET), Peshawar in 2011 and his Master's degree in Electrical Engineering from UET, Peshawar in 2015. He is also pursuing his Ph.D. Currently; he is working as a Lecturer at Department of Electrical Engineering, UET Peshawar. His research interests include image processing, computer vision, machine learning and deep learning.

Nasir Saleem

Engr. Nasir Saleem received the B.S degree in Telecommunication Engineering from University of Engineering and Technology, Peshawar-25000, Pakistan in 2008 and M.S degree in Electrical Engineering from CECOS University, Peshawar, Pakistan in 2012. He was a senior Lecturer at the Institute of Engineering and Technology, Gomal University, D.I.Khan-29050, Pakistan. He is now Assistant Professor in Department of Electrical Engineering, Gomal University, Pakistan. His research interests are in area of digital signal processing, speech processing and enhancement.

Elena Verdú

Elena Verdú received her master's and Ph.D. degrees in telecommunications engineering from the University of Valladolid, Spain, in 1999 and 2010, respectively. She is currently an Associate Professor at Universidad Internacional de La Rioja (UNIR) and member of the Research Group "Data Driven Science" of UNIR. For more than 15 years, she has worked on research projects at both national and European levels. Her research has focused on e-learning technologies, intelligent tutoring systems, competitive learning systems, accessibility, data mining and expert systems.

Numan Khurshid

Numan Khurshid obtained his B.Sc. degree in Electrical Engineering from University of Engineering and Technology (UET), Peshawar in 2011 and his Master's degree in Electrical Engineering from National University of Science and Technology (NUST), Islamabad. He recently acquired his doctorate in Artificial Intelligence from the School of Science and Engineering, Lahore University of Management Sciences (LUMS), Lahore. Currently, he is working as a Graduate Research Associate at the National Center of Artificial Intelligence, UET Peshawar. His research interests include image processing, computer vision, natural language processing, machine learning, and deep learning.

# An Empiric Analysis of Wavelet-Based Feature Extraction on Deep Learning and Machine Learning Algorithms for Arrhythmia Classification

Ritu Singh[1]*, Navin Rajpal[1], Rajesh Mehta[2]

[1] University School of Information and Communication Technology, Guru Gobind Singh Indraprastha University, Dwarka, New-Delhi (India)
[2] Computer Science and Engineering Department, Thapar Institute of Engineering and Technology, Patiala, Punjab (India)

Received 15 April 2020 | Accepted 30 October 2020 | Published 11 November 2020

unir
LA UNIVERSIDAD EN INTERNET

## Abstract

The aberration in human electrocardiogram (ECG) affects cardiovascular events that may lead to arrhythmias. Many automation systems for ECG classification exist, but the ambiguity to wisely employ the in-built feature extraction or expert based manual feature extraction before classification still needs recognition. The proposed work compares and presents the enactment of using machine learning and deep learning classification on time series sequences. The two classifiers, namely the Support Vector Machine (SVM) and the Bi-directional Long Short-Term Memory (BiLSTM) network, are separately trained by direct ECG samples and extracted feature vectors using multiresolution analysis of Maximal Overlap Discrete Wavelet Transform (MODWT). Single beat segmentation with R-peaks and QRS detection is also involved with 6 morphological and 12 statistical feature extraction. The two benchmark datasets, multi-class, and binary class, are acquired from the PhysioNet database. For the binary dataset, BiLSTM with direct samples and with feature extraction gives 58.1% and 80.7% testing accuracy, respectively, whereas SVM outperforms with 99.88% accuracy. For the multi-class dataset, BiLSTM classification accuracy with the direct sample and the extracted feature is 49.6% and 95.4%, whereas SVM shows 99.44%. The efficient statistical workout depicts that the extracted feature-based selection of data can deliver distinguished outcomes compared with raw ECG data or in-built automatic feature extraction. The machine learning classifiers like SVM with knowledge-based feature extraction can equally or better perform than Bi-LSTM network for certain datasets.

## Keywords

Arrhythmia Classification, Bi-Long Short-Term Memory, Multi-Resolution Analysis, Support Vector Machine, Wavelet Transform.

DOI: 10.9781/ijimai.2020.11.005

## I. Introduction

THE automation in electrocardiogram (ECG) measurement enables users to monitor their cardiac signals using smart portable devices like wearables [1]. Any heart complexity is immediately observed, reported, or consulted to the experts. With these advancements, ECG classification and analysis are upgraded from machine learning to deep learning. The change of data from 1D to 2D or 3D or vice versa requires high accuracy and low computational time. The computer configuration needs to get compatible with new technologies.

There are two phases for the automatic detection and realization of any cardiac anomaly. These phases are feature extraction and classification, such as binary or multi-class. The feature extraction stage gives flexibility to any algorithm to become efficient and increase the performance rate. It is based on a thorough knowledge of the inputs and dataset. With expert experience added, it becomes a powerful tool to extract the desired features easily. If features extracted are large in dimensions or direct data samples are acquired, the need comes from feature compression [2] or reduction. This feature selection filters primary significant features that make an easy input for classifiers. The second stage is classification, where the classifier algorithm gets trained by the collected input feature dataset to predict the test data and unknown data. This type of automation is seen in traditional models that use artificial intelligence and machine learning. The traditional models require a separate feature extraction module like features extracted by experience, signal processing techniques, and classification algorithms. These may include wavelet features [3], [4], [5], Principal Component Analysis (PCA) [6], Independent Component Analysis (ICA) [7], [8], and statistical features [9]. Wavelet Transform (WT) has shown a high impact on ECG analysis as wavelet decomposition gives its sub-bands and coefficients at different levels. This disintegration helps in finding unique features for analysis. A wavelet design devoted to noise suppression with the Hidden Markov Model (HMM) gives successful multi-classification with distinctive feature extraction [10].

* Corresponding author.
E-mail address: ritu.usict.041164@ipu.ac.in

Recently, technology up-gradation has given deep learning algorithms that have a single end-to-end structure for feature extraction and classification. These innovations have given many new classification algorithms like Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN) [11], [12], [13], a hybrid structure like CNN with Bidirectional LSTM [14] and active classification using deep learning networks [15]. There is another interesting combination of CNN and LSTM that feature extract and classify ECG signals of variable length and achieving accuracy of 98.10% [16]. These models learn features automatically and get trained.

This experimental study, analyze and compare BiLSTM network and SVM classification algorithm on 1D sequential ECG data. The paper contributes towards,

- Implementing discrete wavelet-based denoising and Maximal Overlap Discrete Wavelet Transform (MODWT) based feature extraction method for extracting 6 morphological and 12 statistical ECG attributes.
- Providing no information loss due to time in-variant, non-orthogonal, less variable estimation, and stationary detail time series achieved by the multi-resolution analysis of MODWT.
- Illustrating the application and the data-based choice to use machine learning or deep learning for 1-D signals of arbitrary length.
- Conduction of a systematic experiment that demonstrates that SVM can perform as good as the BiLSTM network on the same benchmark PhysioNet ECG datasets in similar conditions.

In addition to this, the arrhythmic features are discussed and supervised by cardiac experts. The classification outcome shows that extracted featured ECG data yields higher performance than raw ECG data for deep learning and machine learning classification techniques.

## II. Preliminaries

### A. Multi-resolution Wavelet Transform

Wavelet Transform (WT) has a wide application area for non-stationary electrical signals like biomedical. WT provides time-frequency information simultaneously. The signal representation at various frequency levels and analyzing it through high and low pass filters at different scales give the concept of multi-resolution analysis. MODWT is indifferent to the start point selection of a time series sequence. MODWT implements DWT twice, once to original series and another to its transformation, and then merges the outputs. MODWT coefficients are scaling ($\sim s_{k,m}$), wavelet ($\sim w_{k,m}$), approximation ($\sim a_{k,m}$) and detail ($\sim d_{k,m}$). These coefficients are described as,

$$\sim s_{k,m} = \sum_{l=0}^{L_k-1} \sim g_{k,l}\, x_{m-1 \bmod N} \quad (1)$$

$$\sim w_{k,m} = \sum_{l=0}^{L_k-1} \sim h_{k,l}\, x_{m-1 \bmod N} \quad (2)$$

$$\sim a_{k,m} = \sum_{l=0}^{N-1} \sim g^{o}_{k,l}\, \sim s_{k,m+1 \bmod N} \quad (3)$$

$$\sim d_{k,m} = \sum_{l=0}^{N-1} \sim h^{o}_{k,l}\, \sim w_{k,m+1 \bmod N} \quad (4)$$

where $\sim g^{o} = \sim g$, periodized to length N and $\sim h^{o} = \sim h$, periodized to length N [17].

MODWT can manage arbitrary sample dimensions as it is an undecimated type of wavelet transform. The multi-resolution of MODWT exhibits the zero-phase filtering giving an advantage to the extracted features to be time-aligned. The characteristics like less variable estimation and content retention help MODWT be well-suited with time series as recommended in [18], [19].

### B. Support Vector Machine (SVM)

SVM represents supervised machine learning models implementing kernel functions for non-linear mapping space. SVM can handle binary and multi-class problems efficiently. Many real-world applications are successfully implemented using support vector classification. The working is based on an optimal separable hyperplane [20]. The hyperplane corresponds to a non-linear decision margin for classification.

SVM deals with noisy and sparse datasets efficiently. SVM is an exception in handling large and small datasets.

### C. Bidirectional Long Short-Memory (BiLSTM) Network

After the growth of machine learning, RNN has ideally started by retaining and utilizing state information. Storing previous time information leads to a memory unit. An improvement over RNN, i.e., LSTM classifier has a gating mechanism that manages long term input data. It has three layers: input, forget, and output layer. For a complete long sequence of data, Bidirectional RNN proposes forward and backward state RNN.

BiLSTM network uses two LSTMs for both the past token state and future token state. The information is processed from left to right and vice-versa. For each time stride, there is a hidden forward layer containing an unknown unit function that operates on the previous hidden state, input forward state, and hidden back layer having a hidden unit that stores future hidden state and input to the current step. A long vector comprises forward and backward representation. Moreover, the final outputs are the predictions [21].

TABLE I. Dataset Acquisition from PhysioNet

| PhysioNet Datasets | Description | Size of ECG Signal |
|---|---|---|
| **For Binary**<br>DB1:<br>the PhysioNet 2017 Challenge<br>Sampling rate:<br>300 Hz at 16-bit resolution | Normal Signal (N_S)<br><br>Atrial Fibrillati-on Signal (AFib_S) | **Total:**<br>**5665 x 9000**<br>4937 x 9000<br>718 x 9000 |
| **For Multi-class**<br>DB2: includes 3 Datasets<br>• MIT-BIH Arrhythmia<br>• The BIDMC Congestive Heart Failure<br>• MIT-BIH Normal Sinus Rhythm<br>Sampling rate :<br>128 Hz at 16-bit resolution | Arrhythmia Signal (A_S)<br><br>Congestive heart failure (CHF_S)<br><br>Normal Sinus (NS_S) | **Total:**<br>**162 x 65536**<br>96 x 65536<br><br>30 x 65536<br><br>36 x 65536 |

## III. Proposed MODWT Multiresolution Analysis Based SVM and BiLSTM Scheme

The detailed feature extraction and classification modules are structured in Fig. 1.

### A. ECG Dataset Acquisition

The frequently used PhysioNet databases are involved in the present study. A detailed description of the dataset acquisition is tabulated in Table I. For the binary dataset, the PhysioNet 2017 Challenge [22] includes two types of ECG signals, such as Normal (N_S) and Atrial fibrillation (AFib_S). The data is stored at 300 Hz with 0.5-40 Hz of bandwidth. The direct samples of each signal give accurate signal statistics. The length of each signal is trimmed to 9000 samples for balanced data collection. The multi-class dataset requires three different ECG signals from three different PhysioNet databases, namely MIT-BIH Arrhythmia Database for Arrhythmia ECG Signal

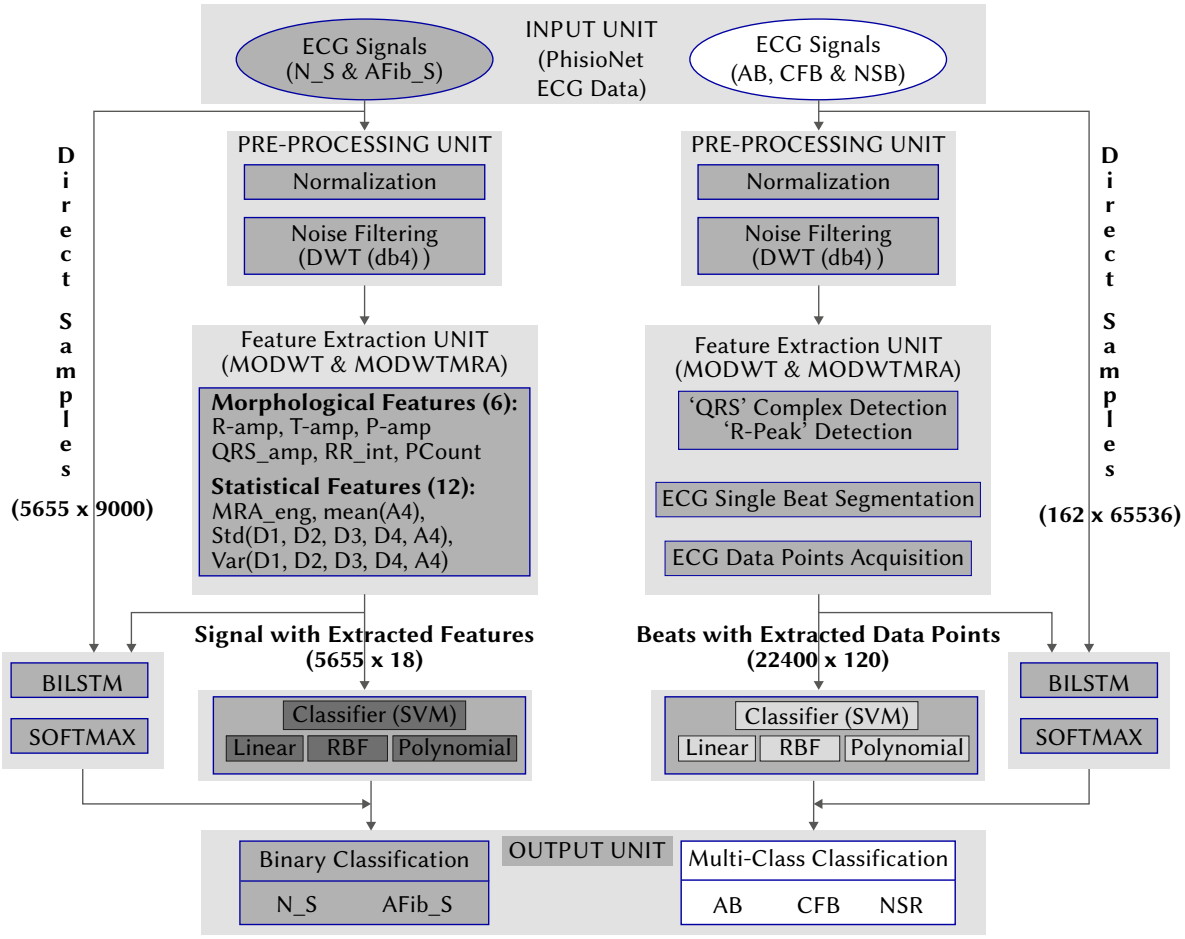



Fig. 1. Block Diagram of proposed feature extraction and classification module.

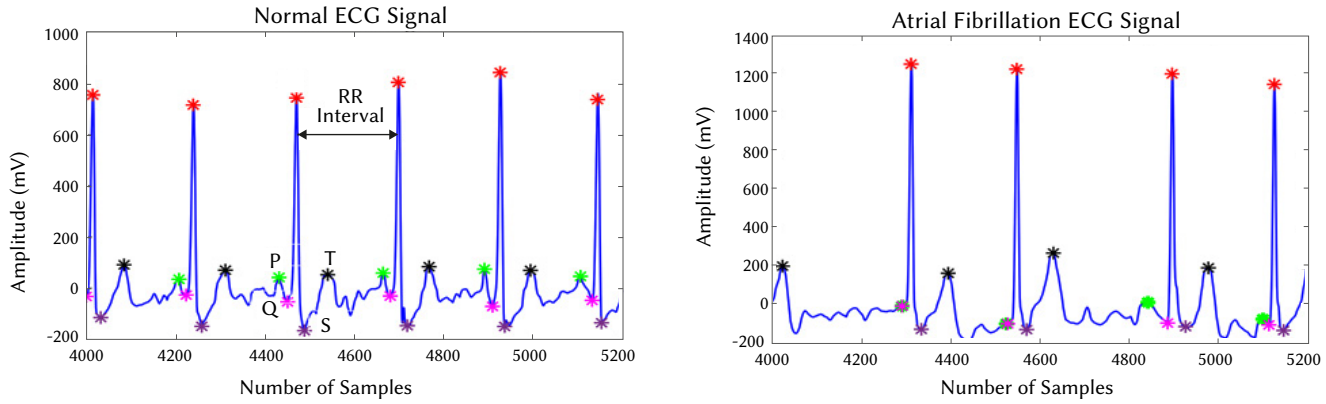



Fig. 2. Extracted ECG features of DB1 for binary classification.

(A_S), the BIDMC Congestive Heart Failure Database for Congestive heart failure Signal (CHF_S) and MIT-BIH Normal Sinus Rhythm Database for Normal Sinus Signal (NS_S). The data collection has 65536 samples of each ECG recording, which is sampled at 128 Hz [23].

### B. Pre-processing Unit

During the pre-processing stage, the collection of raw ECG samples is refined by two processes, such as normalization that returns data with the centre to zero and standard deviation to one. The amplitude variation is reduced to a minimum, and consistent data is available for further processing. The next step is to filter ECG and remove noise artifacts like baseline wander and power line interferences. In the present work, the discrete wavelet transform (DWT) is implemented using the Daubechies wavelet family (db4). The wavelet decomposition, removal of undesired detail, and approximate coefficient and reconstruction of signal results in filtered ECG signal [24]. Fig. 2 and Fig. 3 displays normalized and filtered ECG signal.

### C. Feature Extraction

For the feature extraction process, a preliminary session was conducted to determine the difference between arrhythmic conditions involved in the present study. Cardiac experts supervise the feature recognition workout. MODWT and MODWT Multiresolution Analysis (MODWTMRA) are applied for extracting the distinctive attributes. The filtered ECG signal is decomposed to level 4 using Daubechies(db4) wavelet, and MRA is applied that results in detail

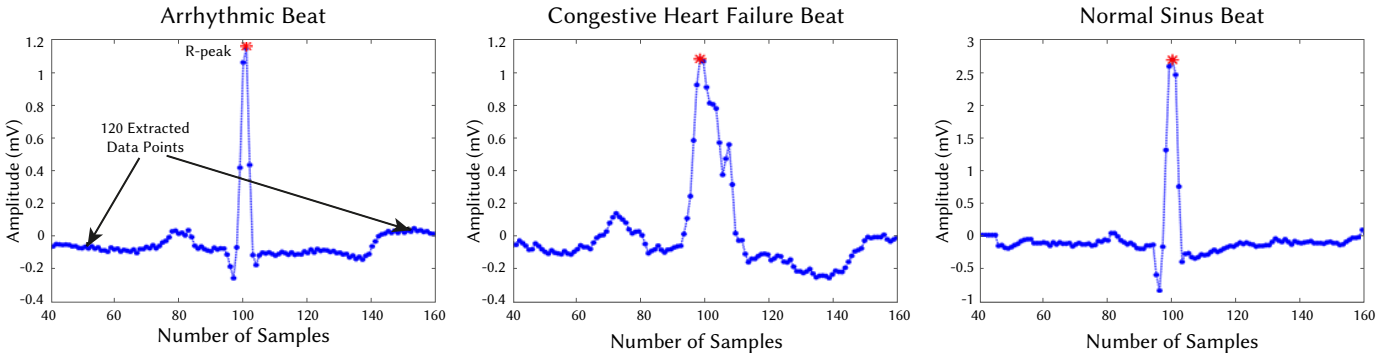



Fig. 3. Beat segmentation of ECG signals (DB2) for multi-class classification.

(D1, D2, D3, D4) and an approximation coefficient (A4). D4 exactly matches the original sample coordinates. So, it is used for extracting the morphological features using signal processing techniques [25].

For the binary dataset DB1, 18 feature vectors comprising 6 morphological and 12 statistical features are extracted. The morphological features are the amplitudes of prominent peaks (P, R, T,'QRS' complex), RR interval, and Pcount. As it is observed in Fig. 2 that in atrial fibrillation P peaks are not prominent, and their count varies from normal ECG. Also, there is a difference in RR interval, Ramp, and Tamp. The statistics are applied to the coefficients reducing their dimensions to achieve better results. The attributes are mean of A4, standard deviation, and variance of D1, D2, D3, D4, and A4. And lastly, maximum MRA energy from all scales. The ECG signal dimension reduces from 5655 x 9000 to 5665 x 18 to be used by the classifier.

For the multi-class dataset (DB2), the ECG signal count is few for classification. So, beat segmentation from 162 ECG signals is required. The beat segmentation requires R peak location and 99 samples before R peak and 100 samples after R peak, comprising 120 samples for each heartbeat count. It is observed that the three different ECG signals such as A_S, CHF_S, and NS_S are very similar in morphological metrics, and only the slope and QRS width have shown variation, as presented in Fig. 3. So, these extracted 120 data points of every single heartbeat can directly be used. The ECG signal dimensions reduce from 162 x 65536 to 22400 x 120 ECG beats and can be used by the classifier.

### *D. Classification*

The differentiating feature vectors of datasets DB1 and DB2 are inputted to the classifiers such as the BiLSTM network and SVM. The two categories of data are imported to a simple BiLSTM network layer. For DB1, the input to BiLSTM is direct samples (5665 x 9000) and featured data (5665 x 18). For DB2, the input to BiLSTM is direct samples (162 x 65536) and featured data (22400 x 120). The output size of the BiLSTM layer is kept 100 units, and the output mode is set to 'last' that maps input signal into 100 features. The other attributes of BiLSTM training are adaptive moment estimation, mini-batch size of 150 for each epoch, maximum epochs of 10, Initial learning rate as 0.01, and gradient threshold is set to 1 to stabilize output.

In parallel, SVM is also used as a classifier, and the input is 5665 x 18 featured ECG signals, and 22400 x 120 featured ECG beats. SVM uses three kernel functions that are linear, rbf, and quadratic or polynomial.

## IV. Experimental Results

The proposed classification setup requires both the ECG signal as well as ECG beat. So, features are extracted, and beats are detected from the signal. For extensive performance analysis and evaluation, two different datasets are created from PhysioNet, namely DB1(binary-dataset) comprising normal(N_S) and abnormal (AFib_S) signals, and DB2(multi-class) comprising three different ECG beats such as AB, CFB, and NSB. The classification results are realized using MATLAB (R2018 working environment for academic use), and NVIDIA Discrete graphics with GPU are used for the training process.

ECG data signals and beats are grouped as testing and training data. The training process helps the classifier train on existing data, whereas the testing process checks the accuracy of the classifier on unknown or new data. As for DB1, the AFib_S signals are very few compared to N_S (718: 4937), so data augmentation is proposed that is also known as oversampling. The MATLAB function 'repmat' is used for this purpose. As for DB2, the three different ECG beats are good in the count. So, there is no need of data repetition. The data partitioning scheme is not required for the BiLSTM network as the neural network shuffles the data automatically. Nevertheless, for SVM, 5-fold and 10-fold cross-validation schemes are implemented for DB1 and DB2, respectively. The proposed testing and training arrangement yield efficient results. Table IV gives training and testing of data information.

Fig. 4 to Fig. 7 show the accuracy obtained with the BiLSTM network scope. Each plot is divided into two sections. The top section depicts the training process, and the bottom section depicts the training loss simultaneously. The respective confusion matrix is also shown. Fig. 4 presents the classification through the BiLSTM network for DB1 using direct ECG samples showing training and testing accuracy of 61.6% and 58.1%, respectively. Moreover, the same network inputted with a featured dataset, as shown in Fig. 5, depicts an improvement of training and testing accuracy of 81.5% and 80.7%, respectively. In the case of DB2, Fig. 6 shows the BiLSTM network with direct ECG samples, and Fig. 7 shows a vast improvement in training and testing accuracy from 88.8% to 95.9% and 49.6% to 95.4% respectively. Unlike the previous result, 120 segmented ECG data points help in the improvement of accuracy.

The statistical parameters are Overall Accuracy Analysis (OAA), Precision (%), Recall (%) and F1Score that are defined by,

$$OAA(\%) = \frac{\text{TPR+TNR}}{\text{TPR+TNR+FPR+FNR}} \times 100 \tag{5}$$

$$Precision(\%) = \frac{\text{TPR}}{\text{TPR+FPR}} \times 100 \tag{6}$$

$$Recall\ (\%) = \frac{\text{TPR}}{\text{TPR+FNR}} \times 100\,\% \tag{7}$$

$$F1Score\ (\%) = \frac{\text{2*(Recall * Precision)}}{\text{Recall + Precision}} \times 100\,\% \tag{8}$$

where TPR: True Positive Response, FPR: False Positive Response, FNR: False Negative Response, and TNR: True Negative Response. TPR means truly existing and detected signal. FPR means not a true response but detected. FNR means to be a true response but not detected. F1 Score means minimum and maximum optimal recognition. Table II and Table III tabulates the classification performance of binary and multi-class SVM.

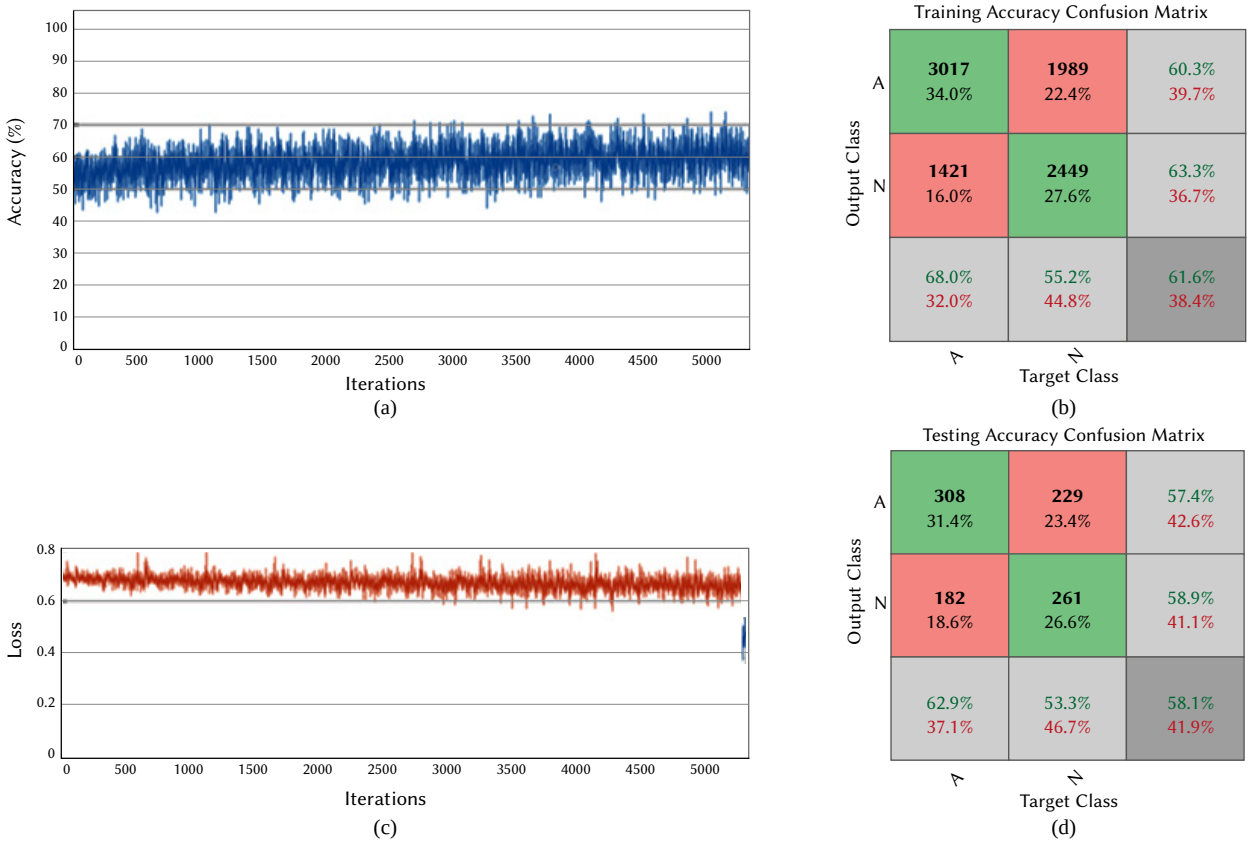



Fig. 4. BiLSTM performance on direct ECG samples (DB1): (a) Training accuracy with iterations; (b) Training Confusion matrix; (c) Loss with iterations; (d) Testing Confusion matrix.

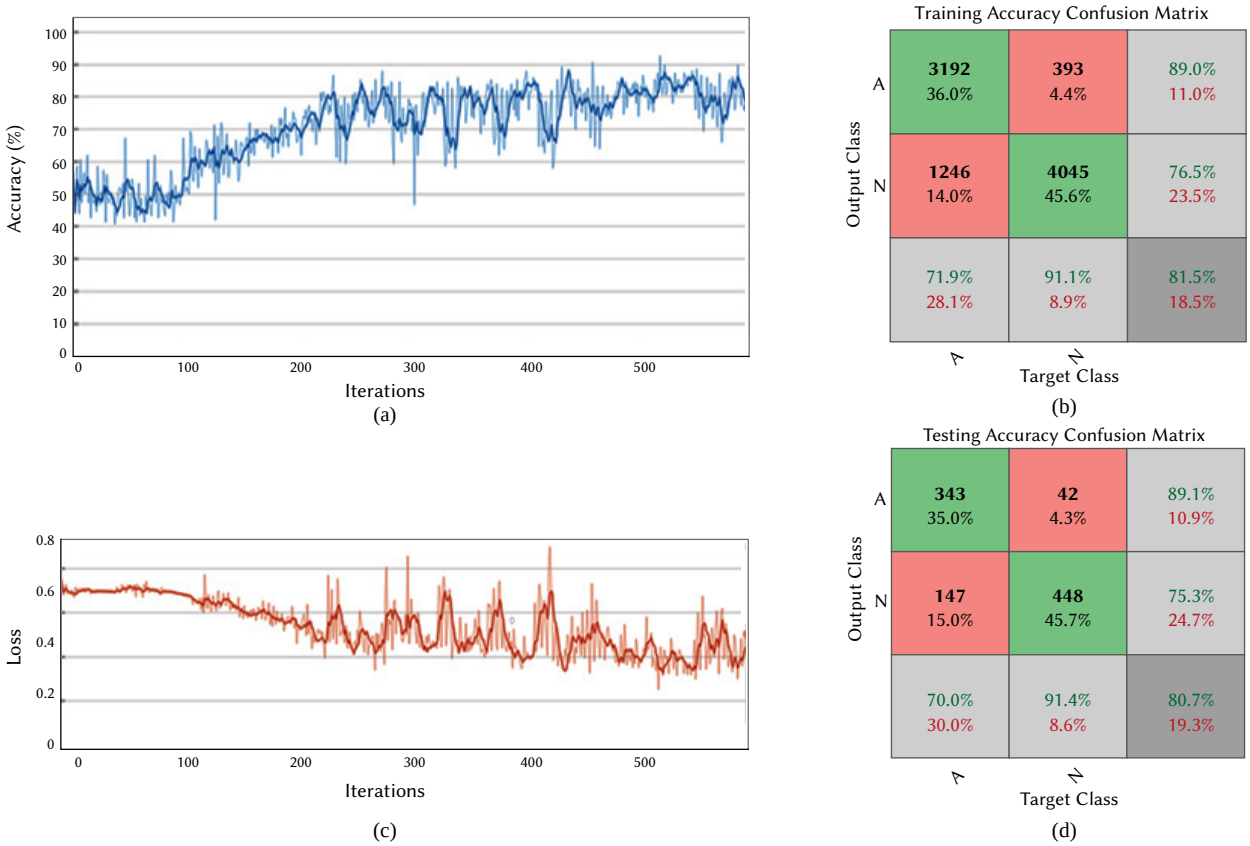
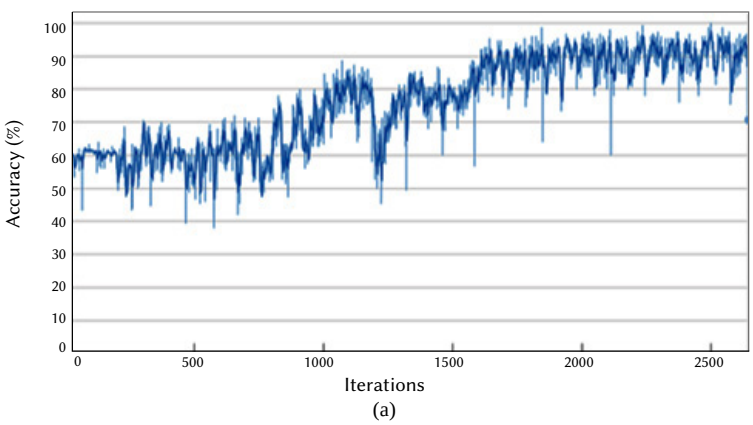
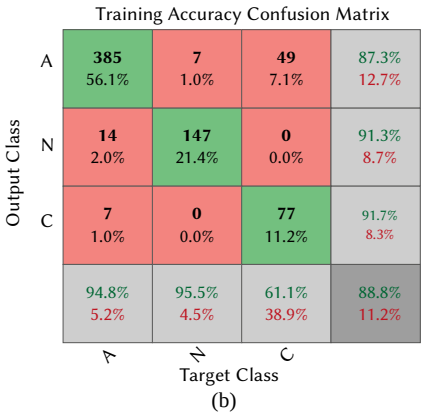


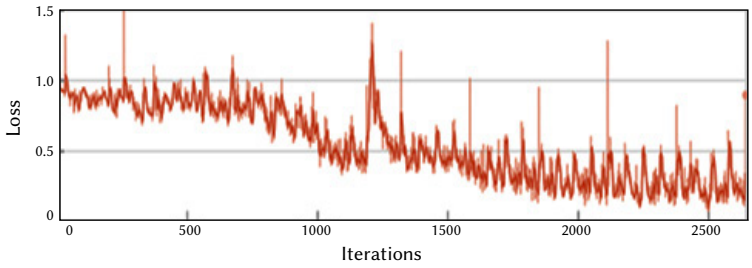

Fig. 5. BiLSTM performance on extracted features of ECG samples (DB1): (a) Training accuracy with iterations; (b) Training Confusion matrix.; (c) Loss with iterations; (d) Testing Confusion matrix.



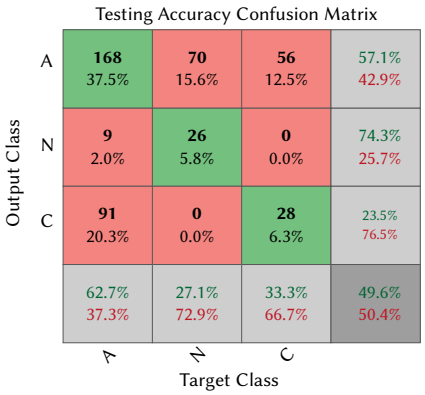

(a)




(b)

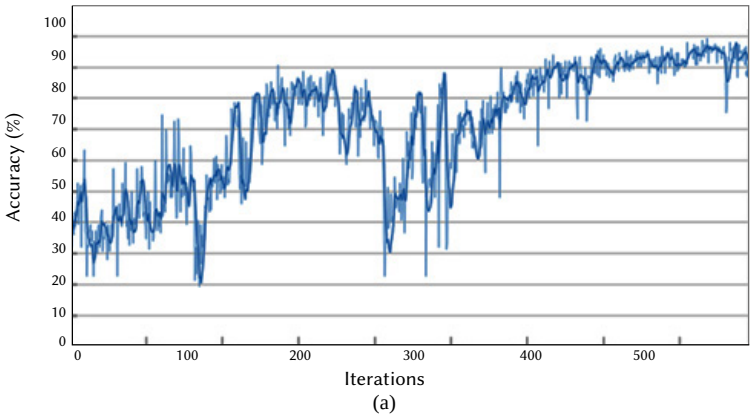



(c)

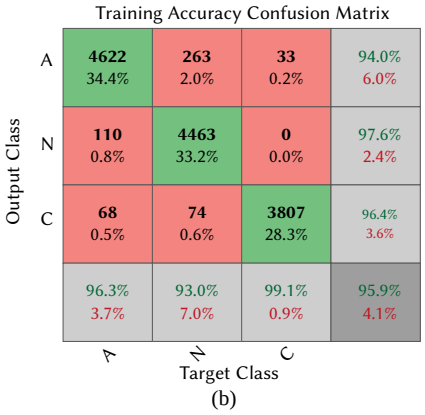



(b)

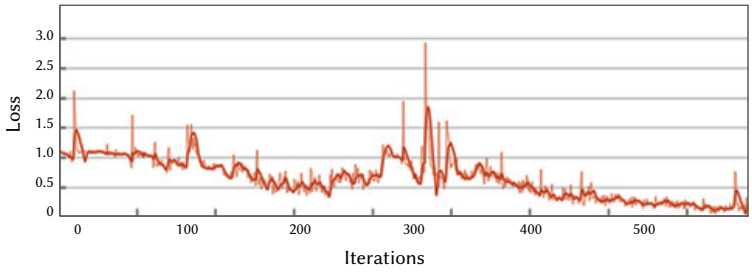
Fig. 6. BiLSTM performance on direct ECG samples (DB2): (a) Training accuracy with iterations; (b) Training Confusion matrix.; (c) Loss with iterations; (d) Testing Confusion matrix.



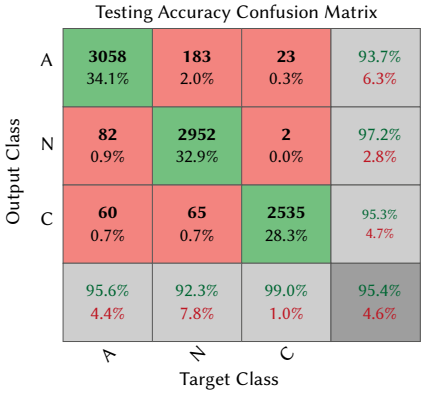

(a)




(b)




(c)




(b)

Fig. 7. BiLSTM performance on segmented ECG beats (DB2): (a) Training accuracy with iterations; (b) Training Confusion matrix.; (c) Loss with iterations; (d) Testing Confusion matrix.

TABLE II. Binary Classification Outcomes of SVM for DB1

| Classifier Type | Confusion Matrix | | Precision, % | | Recall, % | | F1 Score, % | | OAA, % |
|---|---|---|---|---|---|---|---|---|---|
| | N_S | AFib_S | N_S | AFib_S | N_S | AFib_S | N_S | AFib_S | |
| Linear | 4789 | 237 | 99.66 | 95.40 | 95.28 | 99.67 | 97.42 | 97.49 | 97.46 |
| | 16 | 4921 | | | | | | | |
| RBF | 5026 | 0 | 99.74 | 100 | 100 | 99.73 | 99.87 | 99.86 | 99.87 |
| | 13 | 4924 | | | | | | | |
| Polynomial | 5026 | 0 | 99.76 | 100 | 100 | 99.75 | 99.88 | 99.87 | 99.88 |
| | 12 | 4925 | | | | | | | |

TABLE III. Multi-class Classification Outcomes of SVM for DB2

| Classifier Type | Confusion Matrix | | | Precision, % | | | Recall, % | | | F1 Score, % | | | OAA, % |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AB | CFB | NSB | AB | CFB | NSB | AB | CFB | NSB | AB | CFB | NSB | |
| Linear | 7984 | 13 | 3 | 99.57 | 98.18 | 98.83 | 99.8 | 98.15 | 98.63 | 99.68 | 98.17 | 98.73 | 98.92 |
| | 28 | 6282 | 90 | | | | | | | | | | |
| | 6 | 103 | 7891 | | | | | | | | | | |
| RBF | 7995 | 3 | 2 | 99.98 | 98.39 | 99.73 | 99.93 | 99.68 | 98.73 | 99.96 | 99.03 | 99.23 | 99.44 |
| | 1 | 6380 | 19 | | | | | | | | | | |
| | 0 | 101 | 7899 | | | | | | | | | | |
| Polynomial | 8000 | 0 | 0 | 99.67 | 99.03 | 99.34 | 100 | 98.87 | 99.12 | 99.82 | 98.95 | 99.23 | 99.37 |
| | 20 | 6328 | 52 | | | | | | | | | | |
| | 8 | 62 | 7930 | | | | | | | | | | |

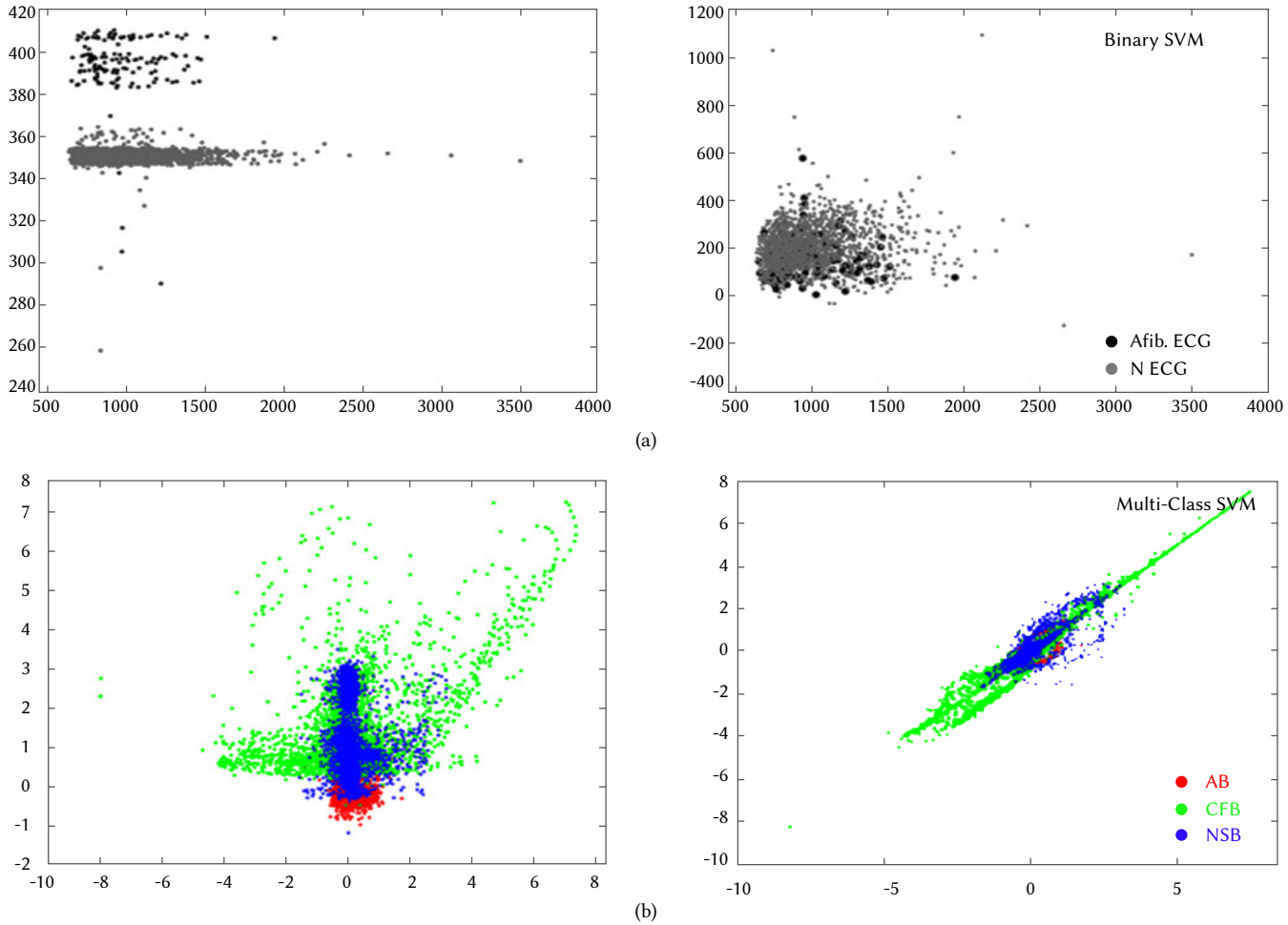



Fig. 8. Scatter diagram for SVM: (a) Binary Classification, (b) Multi-class classification.

Fig. 8 displays a scatter diagram for SVM that discriminates coefficients of binary and multi-class datasets. The experimental results predict that with a large number of beat counts and a large dataset, SVM gives better accuracy for non-linear and non-stationary biological signals like ECG compared to the BiLSTM deep learning network.

## V. Discussion

The impact of employing different classification techniques on direct, in-built, and knowledge-based handcrafted features of binary and multi-class ECG datasets has shown consequential observations, as indicated in Table IV.

The feature extraction before applying classification shows much better performance in the present study, and the same is also reported in [34]. For both the datasets, the accuracy rate of 95% and above is achieved only in the knowledge-based extracted features of ECG signals. The statistical variations can be justified by the points described below.

### *A. ECG Feature Set*

In the case of a binary dataset, using knowledge-based 18 extracted attributes with the BiLSTM network results in an increase of 19.9 % training accuracy and 22.6 % of testing accuracy compared to using direct raw ECG samples.

The same feature set with SVM results in an increase of 19.18 % performance accuracy compared with the BiLSTM network. This means that if known features of arrhythmic ECG signal are differentiated and extracted, as shown in Fig. 2, machine learning can perform better than deep learning in such cases.

Similarly, for a multi-class dataset, ECG beat segmentation is done to demonstrate another positive impact of extracting PQRST data points of a single beat. These are hand-crafted direct 120 ECG data points of each heartbeat, as shown in Fig. 3. Using these features with the BiLSTM network results in an increase of 7.1 % training accuracy and 45.8 % of testing accuracy compared with using direct raw ECG samples or whole signal as input. The same feature set with SVM results in an increase of only 4.04 % accuracy compared with the BiLSTM network. This illustrates that instead of using all direct raw ECG samples, it is beneficial to use required and informative features with deep learning to increase performance accuracy above 95%. Also, machine learning algorithms like SVM can perform equal or better than deep learning networks like BiLSTM.

TABLE IV. Performance Comparison of the Implemented SVM and BiLSTM Models

| Classification | Data Partitions | Feature set | Classifier | | Accuracy (%) |
|---|---|---|---|---|---|
| **Binary No. of classes: 2** | Training Data: 8876<br>Testing Data: 980<br><br>For SVM:<br>5-fold cross validation | Direct Samples | BiLSTM | Training | 61.6 |
| | | | | Testing | 58.1 |
| | | MODWT & MODWT MRA based morphological and statistical features (18 features) | BiLSTM | Training | 81.5 |
| | | | | Testing | 80.7 |
| | | | SVM | Linear | 97.46 |
| | | | | RBF | 99.87 |
| | | | | Polynomial | 99.88 |
| **Multi-class No. of classes: 3** | Training Data: 13340<br>Testing Data: 8960<br><br>For SVM:<br>10-fold cross validation | Direct Samples | BiLSTM | Training | 88.8 |
| | | | | Testing | 49.6 |
| | | MODWT & MODWT MRA based Beat Segmentation (120 Data points) | BiLSTM | Training | 95.9 |
| | | | | Testing | 95.4 |
| | | | SVM | Linear | 98.92 |
| | | | | RBF | 99.44 |
| | | | | Polynomial | 99.37 |

TABLE V. Performance Comparison of the Proposed Models with Other State-of-the-art Methods

| Literature | Classes | Number of ECG beats | Extracted Features | Classifier | Accuracy (%) |
|---|---|---|---|---|---|
| **Sahoo et al. (2017)** [26] | 4 | 1071 | MRA of DWT (Temporal and morphological) | SVM | 98.39 % |
| **Plawiak (2018)** [27] | 17 | 1000 | Genetic optimization, selection and the spectral power density estimation | SVM | 98.85% |
| **Guerra et al. (2019)** [28] | 4 | 49,691 | Wavelets, Higher order statistics, morphological and local binary patterns | Multiple SVM combination | 94.50% |
| **Zubair et al. (2016)** [29] | 5 | - | End-to-end | CNN | 92.70% |
| **Acharya et al. (2017)** [30] | 2 | 110094 | End-to-end | CNN | 95.22% |
| **Acharya et al. (2017)** [31] | 5 | 109,449 | End-to-end | CNN | 94.03% |
| **Lodhi et al. (2018)** [32] | 2 | 81,652 | End-to-end | CNN | 93.53% |
| **Lui et al. (2018)** [33] | 4 | - | End-to-end | CNN-LSTM | 94.62% |
| **Proposed models (2020)** | 2 | 186,615 | Direct Samples | BiLSTM | 58.1% |
| | | | MODWT & MODWT MRA based features | BiLSTM | 80.7% |
| | | | | SVM | 99.88% |
| | 3 | 22,400 | Direct Samples | BiLSTM | 49.6% |
| | | | MODWT & MODWT MRA based beat segmentation | BiLSTM | 95.4% |
| | | | | **SVM** | **99.44%** |

### B. Performance Comparison with Existing Literatures

The efficient classification outcomes performed by different methods recently are illustrated in Table V. The robust feature extraction techniques like wavelet decomposition are used before classifiers like SVM, as reported in [26], [28]. Sahoo et al. [26] detected the QRS complex using MRA of WT with SVM classification on MIT–BIH ECG database of PhysioNet achieving 98.39% accuracy and a meager error rate 0.42%. In 2018, Pawel Pławiak achieved 98.85% accuracy on ECG fragments using feature extraction with pre-processing. ECG characteristics were estimated using PSD and tested using genetic optimization and selection before employing SVM classification on 1000 cardiac beats [27]. An ensemble SVM, i.e., multi SVM approach, is demonstrated with wavelet-based, HOS, LBP, and many amplitude values for feature extraction with specific SVMs [28]. The ensemble methodology implemented showed satisfactory performance of 94.50 % of accuracy.

The automatic in-built feature extraction concept is also known as the End-to-end technique, is used in deep learning algorithms, as reported in [29]-[33]. Zubair et al. [29] employed a small patient-specific ECG dataset to implement CNN achieving classification accuracy of 92.50 % for five different beats. Acharya et al. [30] proposed CNN to diagnose normal and myocardial beat with an accuracy of 95.22%. They investigated ECG beats with and without noise removed. Another CNN model was designed by Acharya et al. [31] in 2017, depicting 94.03% accuracy with high-frequency noise removal technique on 109,449 ECG beats. They classified five different ECG classes with improved generalization capability. Lodhi et al. [32] achieved 93.53 % accuracy by designing a 20-layered CNN model for binary classification, including 81,652 beats. Another model introduced by Lui et al. [33] has a sequence of CNN and BiLSTM for multi-class MI diagnosis classifying 4 categories and achieving a performance rate of 94.62%.

The accuracy of 80.7% achieved by the proposed BiLSTM networks using hand-crafted feature extraction, yet it is lower than the accuracy of 95.4 % achieved by proposed BiLSTM network using informative beat segmented direct ECG data points. Besides, the proposed SVM with MODWT extracted features outperforms CNN and BiLSTM networks with built-in or hand-crafted features by achieving an accuracy rate of 99.88% for binary and 99.44% for multi-classification respectively. More evidence is reported in [35] where the combination of MRA of DWT with Online Sequential Extreme Learning Machine (OSELM) as classifier has achieved a 99.44% accuracy rate for two classes and 98.51% accuracy rate for multi-class, respectively.

### C. Limitations

In the present study, there is the usage of data augmentation for BiLSTM networks, 5-fold, and 10-fold cross-validation for SVM due to small sample size constraints. So, overfitting issues can exist. This limitation can be rectified by experimenting with large size datasets. Moreover, by using same datasets of different studies and same validation methods the results can be directly compared considering similar environment.

## VI. Conclusion

The proposed work is an experimental research analyzing the classification capability using in-built feature extraction of deep learning with machine learning using distinctive knowledge-based feature extraction on time series sequential ECG data. BiLSTM network with automatic feature extraction is implemented on the publicly accessible and available PhysioNet 2017 Challenge dataset, and then the same two-class dataset is treated with SVM using manual feature extraction derived using MODWT, and MODWTMRA. The 18 feature vectors of normal and Atrial Fibrillation ECG signals are extracted under the supervision of cardiac experts. Another dataset comprising of three different classes from the PhysioNet database is also used. For this, feature extraction involves beat segmentation comprising 120 informative data points of each category of ECG beat. In both cases, under similar experimental scenarios, the raw ECG data is firstly fed to BiLSTM networks, then hand-crafted ECG features to the BiLSTM network and SVM. The research outcomes suggest that deep learning with in-built feature extraction cannot always be an efficient method for all types of ECG datasets. However, machine learning with manual feature extraction can prove to show better performance in certain experimental conditions.

The pre-processing and feature extraction are two significant preliminaries before classification for one-dimensional data. The hand-crafted feature extraction involves expert experiences and control of signal data. It is observed that for a long duration dataset instead of training BiLSTM with raw ECG samples, it is justified to train with informative segmented beat data points or distinctive vital feature set for desired outcomes. Also, the appropriate feature extraction like wavelet decomposition can be incorporated in the deep learning algorithms to achieve high-performance classification.

For future direction, the featured input data can be made robust and refined to achieve higher accuracy using network classifiers by applying dimensionality reduction techniques.

## Acknowledgment

The authors would like to express their special gratitude to Dr. Aditya Batra, M.D., D.M (Cardiology), Holy Heart Hospital, Rohtak, Haryana, India and Dr. S.K. Gulati M.D.(Medicine), Bharat Nursing Home, Rohtak, Haryana , India and Dr. C.V. Singh, M.D. D.A.(anesthesiology), New Janta Clinic and Vidya Vision Pathology Centre, Rohtak, Haryana, India for their expert opinions and suggestions for the feature extraction of ECG data.

## References

[1] J. Jeppesen, Jesper, et al. "O-45 Automated Seizure Detection for Epilepsy Patients Using Wearable ECG-Device," *Clinical Neurophysiology*, Elsevier, vol. 130, no. 7, p. e36, 2019, doi: 10.1016/j. clinph.2019.04.360.

[2] S. Chandra, A. Sharma, G.K. Singh, "A Comparative Analysis of Performance of Several Wavelet Based ECG Data Compression Methodologies." *IRBM*, 2020, doi: 10.1016/j.irbm.2020.05.004.

[3] H. Li, D.Yuan , X. Ma , D. Cui , L. Cao L. ,"Genetic Algorithm for the Optimization of Features and Neural Networks in ECG Signals Classification." *Scientific Reports*, Nature Publishing Group, vol. 7, p. 41011, 2017, doi: 10.1038/srep41011.

[4] M.K. Islam, et al. "Study and Analysis of Ecg Signal Using Matlab &labview as Effective Tools." *International Journal of Computer and Electrical Engineering*, , IACSIT Press, vol. 4, no. 3, p. 404, 2012, doi: 10.7763/IJCEE.2012.V4.522.

[5] M. Thomas, et al. "Automatic ECG Arrhythmia Classification Using Dual Tree Complex Wavelet Based Features." *AEU-International Journal of Electronics and Communications*, Elsevier, vol. 69, no. 4, pp. 715–21, 2015, doi: 10.1016/j.aeue.2014.12.013.

[6] F. A. Elhaj, et al. "Arrhythmia Recognition and Classification Using Combined Linear and Nonlinear Features of ECG Signals." *Computer Methods and Programs in Biomedicine*, Elsevier, vol. 127, pp. 52–63, 2016,doi : 10.1016/j.cmpb.2015.12.024.

[7] R. J. Martis, et al. "ECG Beat Classification Using PCA, LDA, ICA and Discrete Wavelet Transform." *Biomedical Signal Processing and Control*, Elsevier. 8, no. 5, pp. 437–48, 2013, doi: 10.1016/j.bspc.2013.01.005.

[8] S-N. Yu, and K-T. Chou. "Integration of Independent Component Analysis and Neural Networks for ECG Beat Classification." *Expert Systems with Applications*, Elsevier ,vol. 34, no. 4, pp. 2841–46,2008, doi : 10.1016/jeswa.2007.05.006.

[9] H. M. Rai, et al. "ECG Signal Processing for Abnormalities Detection

Using Multi-Resolution Wavelet Transform and Artificial Neural Network Classifier." *Measurement*, Elsevier, vol. 46, no. 9, pp. 3238–46, 2013, doi: 10.1016/j.measurement.2013.05.021.

[10] A.K. Sangaiah, M. Arumugam, G.B. Bian, "An intelligent learning approach for improving ECG signal classification and arrhythmia analysis." *Artificial Intelligence in Medicine*, vol. 103, p.101788,2020, doi: 10.1016/j.artmed.2019.101788.

[11] G. Wang, et al. "A Global and Updatable ECG Beat Classification System Based on Recurrent Neural Networks and Active Learning." *Information Sciences*, Elsevier, vol. 501, pp. 523–42, 2019, doi: 10.1016/j.ins.2018.06.062.

[12] B. Hou, et al. "LSTM Based Auto-Encoder Model for ECG Arrhythmias Classification." *IEEE Transactions on Instrumentation and Measurement*, IEEE, 2019. doi: 10.1109/TIM.2019.2910342.

[13] U. B. Baloglu, et al. "Classification of Myocardial Infarction with Multi-Lead ECG Signals and Deep CNN." *Pattern Recognition Letters*, Elsevier, vol. 122, pp. 23–30, 2019, doi: 10.1016/j.patrec.2019.02.016.

[14] F. Zhu, et al. "Electrocardiogram Generation with a Bidirectional LSTM-CNN Generative Adversarial Network." *Scientific Reports*, Nature Publishing Group, vol. 9, no. 1, pp. 1–11, 2019, doi:10.1038/s41598-019-42516-z.

[15] Al Rahhal, et al. "Deep Learning Approach for Active Classification of Electrocardiogram Signals." *Information Sciences*, Elsevier, vol. 345, pp. 340–54,2016, doi: 10.1016/j.ins.2016.01.082.

[16] S. L. Oh, E. Y. Ng, R. S. Tan, U. R. Acharya, "Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats." *Comput Biol Med*, vol. 102, pp. 278–287, 2018, doi: 10.1016/j.compbiomed.2018.06.002.

[17] A.G. Hafez, and E. Ghamry. "Geomagnetic Sudden Commencement Automatic Detection via MODWT." *IEEE Transactions on Geoscience and Remote Sensing*, IEEE, vol. 51, no. 3, pp. 1547–54, 2012, doi: 10.1109/ICCES.2009.5383235.

[18] L. Zhu, Y. Wang, Q. Fan "MODWT-ARMA model for time series prediction." *Applied Mathematical Modelling*, vol. 38, no. 5-6, pp. 1859-65, 2014 , doi :10.1016/j.apm.2013.10.002.

[19] Z. Zhang, Q. K. Telesford, C. Giusti, K.O. Lim, D. S. Bassett, "Choosing wavelet methods, filters, and lengths for functional brain network construction." *PloS one*, vol. 11, no. 6, 2016, doi: 10.1371/journal.pone.0157243.

[20] J. Chorowski, et al. "Review and Performance Comparison of SVM-and ELM-Based Classifiers." *Neurocomputing*, Elsevier, vol. 128, pp. 507–16, 2014, doi: 10.1016/j.neucom.2013.08.009.

[21] T. Chen, R. Xu, Y. He, X. Wang. "Improving Sentiment Analysis via Sentence Type Classification Using BiLSTM-CRF and CNN." *Expert Systems with Applications*, Elsevier, vol. 72, pp. 221–30, 2017, doi: 10.1016/j.eswa.2016.10.065.

[22] G.D. Clifford, et al. "AF Classification from a Short Single Lead ECG Recording: The PhysioNet/Computing in Cardiology Challenge 2017." *Computing in Cardiology (CinC)*, pp. 1–4, 2017, doi: 10.22489/CinC.2017.065-469.

[23] A. L. Goldberger, et al. "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals." *Circulation*, Am Heart Assoc, vol. 101, no. 23, pp. e215--e220,2000, doi: 10.1161/01.CIR.101.23.e215.

[24] R. Singh, R. Mehta and N. Rajpal, "Efficient Wavelet Families for ECG Classification Using Neural Classifiers." *Procedia Computer Science*, Elsevier, vol. 132, pp. 11–21, 2018, doi: 10.1016/j.procs.2018.05.054.

[25] R. Singh, R. Mehta and N. Rajpal, "Wavelet and Kernel Dimensional Reduction on Arrhythmia Classification of ECG Signals." *EAI Endorsed Transactions on Scalable Information Systems: Online First*, EAI, 2020, doi:10.4108/eai.13-7-2018.163095.

[26] S. Sahoo S, B. Kanungo, S. Behera, S. Sabut, "Multiresolution wavelet transform based feature extraction and ECG classification to detect cardiac abnormalities." *Measurement*, vol. 108, pp. 55–66, 2017, doi: 10.1016/j.measurement.2017.05.022.

[27] P. Pławiak, "Novel Methodology of Cardiac Health Recognition Based on ECG Signals and Evolutionary-Neural System." *Expert Systems with Applications*, Elsevier, vol. 92, pp. 334–49, 2018, doi: 10.1016/j.eswa.2017.09.022.

[28] V. Mondéjar-Guerra, et al. "Heartbeat Classification Fusing Temporal and Morphological Information of ECGs via Ensemble of Classifiers." *Biomedical Signal Processing and Control*, Elsevier, vol. 47, pp. 41–48, 2019, doi: 10.1016/j.bspc.2018.08.007.

[29] M. Zubair, et al. "An Automated ECG Beat Classification System Using Convolutional Neural Networks." *2016 6th International Conference on IT Convergence and Security (ICITCS)*, 2016, pp. 1–5, doi: 10.1109/ICITCS.2016.7740310.

[30] U. R. Acharya, et al. "Application of Deep Convolutional Neural Network for Automated Detection of Myocardial Infarction Using ECG Signals." *Information Sciences*, Elsevier, vol. 415, pp. 190–98, 2017, doi: 10.1016/j.ins.2017.06.027.

[31] U. R. Acharya, et al. "A Deep Convolutional Neural Network Model to Classify Heartbeats." *Computers in Biology and Medicine*, Elsevier, vol. 89, pp. 389–96, 2017, doi: 10.1016/j.compbiomed.2017.08.022.

[32] A. M. Lodhi, A.N. Qureshi, U. Sharif, Z. Ashiq, "A Novel Approach Using Voting from ECG Leads to Detect Myocardial Infarction." *Proceedings of SAI Intelligent Systems Conference*, 2018, pp. 337–52, doi: 10.1007/978-3-030-01057-7_27.

[33] H. W. Lui, and K. L. Chow, "Multiclass Classification of Myocardial Infarction with Convolutional and Recurrent Neural Networks for Portable ECG Devices." *Informatics in Medicine Unlocked*, Elsevier, vol. 13, pp. 26–33, 2018, doi: 10.1016/j.imu.2018.08.002.

[34] F. Shaheen, B. Verma, M. Asafuddoula, "Impact of Automatic Feature Extraction in Deep Learning Architecture." *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2016, pp. 1–8, doi:10.1109/DICTA.2016.7797053.

[35] YILDIRIM, "Ecg Beat Detection and Classification System Using Wavelet Transform and Online Sequential Elm." *Journal of Mechanics in Medicine and Biology*, World Scientific, vol. 19, no. 01, p. 1940008, 2019, doi: 10.1142/S0219519419400086.

Ritu Singh

Ritu Singh is a research scholar in Guru Gobind Singh Indraprastha University (GGSIPU). She received her B.E. (Electronics) from Poona University, Maharashtra and M.E (Electronics and Communication) from Maharishi Dayanand University (MDU), Rohtak. She has more than 8 years of teaching experience. Her current research interests include signal processing, soft computing, ECG and machine learning algorithms She has several publications in reputed international journals and conferences.

Navin Rajpal

Navin Rajpal is a Professor at USICT since September 2004. He served as Dean, USICT from October 1, 2011 to September 30, 2014. He completed his BSc (Engineering) in Electronics and Communication from the R.E.C. Kurukshetra, now known as a NIT, Kurukshetra. He completed his MTech and PhD from Computer Science and Engineering Department, IIT, Delhi. He served in various capacities and has more than 31 years of experience in teaching and research. He has supervised several MTech and 12 PhD students. He has published/presented more than 100 research papers in national and international journals/conferences. He is a life member of CSI and ISTE. His areas of interest are computer vision, image processing, pattern recognition, artificial neural networks, computer graphics, algorithms design and digital hardware design.

Rajesh Mehta

Rajesh Mehta is working as an Assistant Professor in Thapar Institute of Engineering and Technology in Computer Science and Engineering Department. He received his PhD in Information Technology from Guru Gobind Singh Indraprastha University (GGSIP) and M. Tech. in Computer Science & Engineering (CSE) from Guru Jambheshwar University, Hisar. He has teaching and research experience of more than of 14 Years. He has published and presented more than 22 papers in SCI indexed journals and International conferences. His current research interests include image processing, signal processing, digital watermarking, machine learning algorithms, genetic algorithm and fuzzy logic.

# Promising Deep Semantic Nuclei Segmentation Models for Multi-Institutional Histopathology Images of Different Organs

Loay Hassan[1,3]*, Adel Saleh[2], Mohamed Abdel-Nasser[1,4], Osama A. Omer[1], Domenec Puig[4]

[1] Electrical Engineering Department, Aswan University (Egypt)
[2] Gaist Solutions Ltd, Skipton BD23 2TZ (UK)
[3] Department of computer science, Arab Academy for Science, Technology and Maritime Transport, Aswan (Egypt)
[4] Department of Computer Engineering and Mathematics, University Rovira i Virgili, 43007 Tarragona (Spain)

Received 24 May 2020 | Accepted 9 October 2020 | Published 27 October 2020



## Abstract

Nuclei segmentation in whole-slide imaging (WSI) plays a crucial role in the field of computational pathology. It is a fundamental task for different applications, such as cancer cell type classification, cancer grading, and cancer subtype classification. However, existing nuclei segmentation methods face many challenges, such as color variation in histopathological images, the overlapping and clumped nuclei, and the ambiguous boundary between different cell nuclei, that limit their performance. In this paper, we present promising deep semantic nuclei segmentation models for multi-institutional WSI images (i.e., collected from different scanners) of different organs. Specifically, we study the performance of pertinent deep learning-based models with nuclei segmentation in WSI images of different stains and various organs. We also propose a feasible deep learning nuclei segmentation model formed by combining robust deep learning architectures. A comprehensive comparative study with existing software and related methods in terms of different evaluation metrics and the number of parameters of each model, emphasizes the efficacy of the proposed nuclei segmentation models.

## Keywords

Digital Pathology, Nuclei Segmentation, Whole Slide Imaging, Deep Learning.

DOI: 10.9781/ijimai.2020.10.004

## I. Introduction

Nowadays, digital pathology is rapidly gaining momentum as a proven and essential technology. Its popularity has grown in the last decade due to the improvements in hardware and software. The whole-slide imaging (WSI) refers to the scanning of conventional glass slides to produce high-resolution digital images slides, that can be stored and accessed using dedicated software. The potential applications of digital pathology comprise cell segmentation, counting cancer cells, and prognosis of cancers.

Cell segmentation refers to the process of identifying groups of pixels that represent cell nuclei. This process is often complicated, especially in the presence of adjacent or overlapping cells and color variation in histopathological images. It is one of the core operations in histopathology image analysis. So, in the context of computational pathology, accurate nuclei segmentation techniques are highly needful for extracting, mining, and interpreting sub-cellular morphologic information from digital slide images. Several extracted descriptors such as cell nuclei shape and number of cell nuclei in WSI images are key components of studies such as the determination of cancer types, cancer grading, and prognosis [1].

Indeed, there is diverse tissue types, variations in staining, and cell types, leading to different visual characteristics of WSI images. These variations make the segmentation of nuclei segmentation a challenging task (see Fig. 1). The visual characteristics of WSI images make it very difficult to develop traditional image processing-based segmentation algorithms that give acceptable nuclei segmentation results. The difficulty increases when the segmentation algorithms handle WSI images taken from several cancer patients and collected at different medical centers for various organs, such as breast, kidney, prostate, and stomach [4]. Existing nuclei segmentation software and toolboxes include Cell profiler [2] and ImageJ-Fiji [3]. Cell Profiler simultaneously measures the size, shape, intensity, and texture of a variety of cell types in a high throughput manner [2]. ImageJ-Fiji exploits the latest software engineering practices to merge powerful software libraries with a wide range of scripting languages to allow fast prototyping of image processing algorithms [3].

In turn, the success of deep learning models with several computer vision-based applications encouraged researchers to make extensive efforts of works attempted at developing image segmentation approaches using deep learning models. Of note, the nuclei segmentation task necessitates an enormous effort to manually create

* Corresponding author.
E-mail address: loaysh2012@gmail.com

pixel-wise annotations to be used for training deep learning models. For instance, a multi-path dilated residual network was proposed in [14] for nuclei segmentation and detection. In [17], a nuclei segmentation method based on deep convolutional neural networks (DCNNs) for histopathology images was proposed. However, existing nuclei segmentation methods may achieve good results with a dataset of WSI images and poor performance with other datasets. The main reasons for these limited results are color variations in histopathological images resulted from acquiring WSI from different scanners, the overlapping and clumped nuclei, and the ambiguous boundary between adjacent cell nuclei.

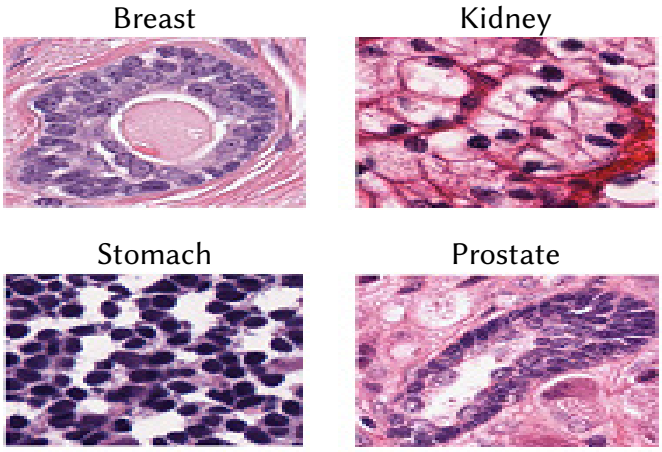



Fig. 1. Examples of WSI images for multi-organ samples.

To tackle these challenges, in this paper, we present promising deep semantic nuclei segmentation models for multi-institutional histopathology images of different organs. These deep learning-based semantic segmentation models are trained in a supervised way to focus on the nuclear regions and to discriminate between nuclear pixels and other pixels. In this way, the models can learn nuclei-aware features, color information, as well as recognizing the complete cells. Indeed, such promising nuclei segmentation models can be used to extract apt features for nuclear morphometrics. Also, it could contribute to the advancement of digital pathology software.

The key contributions of this paper are:

- Study the performance of different deep learning models with nuclei segmentation in WSI images of various stains and various organs. A challenging multi-institutional multi-organ WSI image dataset is used in this paper (publicly available dataset).
- Propose a feasible deep learning nuclei segmentation model formed by combining robust deep learning architectures (so-called PSPSegNet). It achieves 3.48% improvement on the F1-score and 6.62% improvement on aggregated Jaccard index (AJI).
- A comprehensive comparative study with existing software and related methods is presented, in terms of different evaluation metrics and the number of parameters of each model. Also, the use of nuclei segmentation models to count the number of nuclei in WSI images.

Below, we present the remaining sections of this paper. In Section II, we study and discuss the related work. In Section III, we explain the methodology in detail. Section IV includes the experimental results, comparisons and discussion. Section V concludes the paper and gives different points of future work.

## II. Related Work

In the last years, various deep learning models have been employed for performing different segmentation tasks in biology [4]. Generally, most outstanding deep segmentation models are based on convolutional neural networks (CNNs), recurrent neural networks (RNNs), encoder-decoders architecture, and generative adversarial networks (GANs).

Naylor et al. [5] introduced a fully automated method for cell nuclei segmenting in WSI images based on three segmentation models, namely PangNet, a fully convolutional network (FCN), and DeconvNet. They ensembled the three segmentation models, obtaining an F1-score of 0.80. Also, Naylor et al. [6] proposed a segmentation method of nuclei in histopathology data based on CNN. They proposed a distinct idea to segment toughing or overlapping nuclei by formulating the problem as a regression task, where they aim at predicting the distance map of nuclei. They claimed that the main problem in the segmentation of nuclei is that segmentation methods tend to segment adjacent or overlap nuclei one object. Their approach outperforms some related nuclei segmentation methods on the AJI score.

Wang et al. [7] proposed a bending loss regularized network for nuclei segmentation in histopathology images. The proposed bending loss defines high penalties to contour points with large curvatures and applies small penalties to contour points with a slight curvature. Minimizing bending loss can avoid generating contours that encompass multiple nuclei. In the case of histopathology images, nuclei have a smooth shape, and the points on the boundaries of nuclei have small curvature changes. In turn, the points on the contour with large curvature changes have a high probability of being the touching points of two or multiple nuclei. The nuclei segmentation scheme comprises three steps: 1) a preprocessing step for color normalization, 2) an encoder-decoder architecture with the bending loss, and 3) a postprocessing step described in [8] was employed. The proposed model was validated on the MoNuSeg dataset, obtaining an AJI score of 0.621 with the same organ test and score of 0.641 with different organ tests.

Al-Kofahi et al. [9] proposed a three-step cell nuclei segmentation approach: 1) the detection of the cells using a deep learning-based model to obtain pixel probabilities for nuclei, cytoplasm, as well as background, 2) the separation of touching cells based on blob detection and shape-based watershed techniques that can distinguish between the individual nuclei from the nucleus prediction map, and 3) the segmentation of the nucleus and cytoplasm. With four different datasets, they obtained an accuracy of 0.84. Besides, Cui et al. [10] proposed an automatic end-to-end deep neural network algorithm for the segmentation of individual nuclei. They introduced a nucleus-boundary model to predict nuclei and their boundaries simultaneously using a fully convolutional neural network. They obtained the area of each nucleus via a simple, fast, and parameter-free postprocessing procedure. This method can segment a *1000x1000* image in less than 5 seconds, which facilitates precisely segment WSI images in an acceptable time.

In [11], Qu et al. proposed a weakly supervised segmentation framework based on partial points annotation in histopathology images. The framework consists of two stages: 1) a semi-supervised strategy to learn a detection model, and 2) a segmentation model is trained from the detected nuclei locations in a weakly-supervised manner. Specifically, the authors employed the original WSI images and the shape before nuclei to obtain two types of coarse labels from the points annotation using the Voronoi diagram and a k-means clustering algorithm. These rough labels are used to train a deep learning model, and then a dense conditional random field is utilized in the loss function to fine-tune the trained model. With a multi-organ

WSI dataset, they achieved a dice score of 0.73. In [12], a conditional generative adversarial network (cGAN) model was proposed for nuclei segmentation, where the segmentation problem was posed as an image-to-image translation task rather than a classification task. A large dataset of synthetic WSI images with perfect nuclei segmentation labels was generated using an unpaired GAN model. Both synthetic and real data with spectral normalization and gradient penalty for nuclei segmentation were used to train the cGAN model.

Zhou et al. [13] presented a deep learning-based model called contour-aware informative aggregation network (CIA-Net) with a multilevel information aggregation module between two task-specific decoders. Instead of using independent decoders, this model exploits bi-directionally aggregated task-specific features to model the spatial and texture dependencies between nuclei and contour. Besides, a smooth truncated loss is utilized to mitigate the perturbation from outliers. As a result, the CIA-Net model is almost built using informative samples, and so its generalization capability could be enhanced (i.e., with multi-organ multi- center nuclei segmentation tasks). With the 2018 MICCAI challenge of the multi-organ nuclei segmentation dataset, they achieved a Jaccard score of 0.63.

Furthermore, the authors of [14] proposed a multi-path dilated residual network for nuclei segmentation and detection. This network comprises the following: 1) a multi-scale feature extraction step based on D-ResNet and feature pyramid network (FPN), 2) a candidate region network, and 3) a final network for detection and segmentation. The segmentation network involves segmentation, regression, and classification sub-networks. With the MonuSeg dataset, they obtained an AJI of 0.46. Mercadier et al. [15] presented a nuclei segmentation framework based on DCNNs. They formulated the problem as segmentation in a holistic manner rather than the classification of patches. The dataset employed is partially annotated, and they used a weighted background model for the network to give more importance to the boundaries of nuclei.

Furthermore, the authors of [16] used a modified version of the U-Net [28] architecture, so-called U-Net++, in which they combined U-Nets of varying depths. With the nuclei segmentation task, they achieved an improvement of 0.0187 with the intersection-over-union (IoU) metric compared to U-Net. The authors of [17] proposed a nuclei segmentation method based on DCNNs for WSI images. To segment nuclei, they used the Mask R-CNN model [18] with color normalization. In particular, the method includes three major steps: preprocessing, nuclei segmentation, and postprocessing. In the preprocessing step, they applied several augmentation techniques to increase the amount of training data and used a color normalization method to reduce the color variation in WSI images. For nuclei segmentation, they followed the implementation of the Mask R-CNN framework stated in the original paper [18] for the backbone network and employed a feature pyramid network (FPN). In the postprocessing, they applied multiple inference methods to improve the segmentation results, obtaining an F1-score of 0.91.

It is worth noting that color variation in histopathological images, the overlapping and clumped nuclei, and the ambiguous boundary between different cell nuclei limit the performance of the above-mentioned nuclei-segmentation methods. Besides, most of the models proposed for this task are complex and do not give the required results. In this study, we present promising deep semantic nuclei segmentation models for overcoming the above-mentioned limitations. To demonstrate the potency of these models, we consider WSI images collected from different scanners of different organs, namely breast, kidney, colon, stomach, prostate, liver, and bladder.

## III. Methodology

### A. Nuclei Segmentation Framework

Several deep learning-based semantic segmentation approaches have been proposed in the last decade, supported by the outstanding ability of convolutional neural networks (CNN) in producing semantic and hierarchical image features [19]. In our study, we choose five of the most popular models used for semantic segmentation and adapt them to the nuclei cell segmentation task. Fig. 2 presents the framework of nuclei segmentation, which consists of training and testing phases.

As shown in Fig. 2 (a), the training phase includes a preprocessing step, training the segmentation model, and a postprocessing step. In the preprocessing step, we apply the sparse stain color normalization method of [20] to reduce the variability of color between multi-institutional and multi-organs WSI images. The size of WSI images is high (1000×1000 pixels), and thus it very difficult to train deep learning models with such image size. Thus, we split each WSI into four non-overlapped sub-images. In the postprocessing stage, we assemble the segmented masks corresponding to the four non-overlapped sub-images to restore the original image size.

In the training step, we train different deep learning-based segmentation models. To mitigate overfitting and enhance generalization of the deep learning models, we employ data augmentation techniques. Specifically, we randomly crop 200 patches from training images to augment the data. The number of cropped patches is empirically tuned.

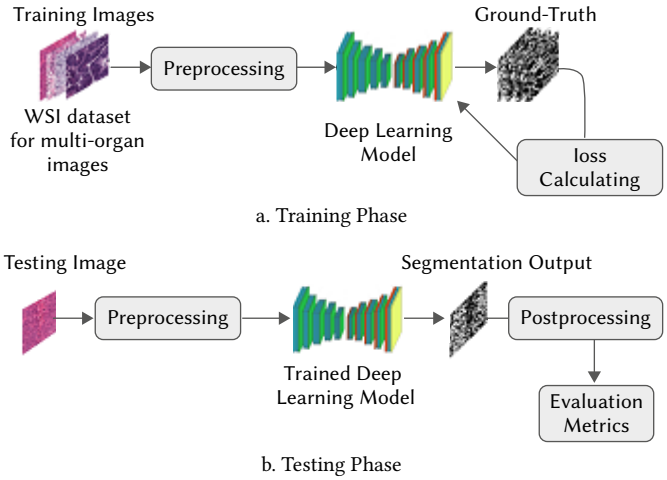



Fig. 2. Nuclei segmentation framework: a) training phase, and b) testing phase.

As shown in Fig. 2 (b), in the test phase we also employ the preprocessing step to normalize the stain of test images and split each test image into four non-overlapped images. Of note, preprocessing step is important to make the same setting used to train the models. After we get the segmented image from a trained model, we apply postprocessing operations to restore the original image size. Besides, we use a connected component algorithm to detect cells and count them. Finally, we evaluate the performance of the nuclei segmentation models in terms of pixel-level F1-score (Dice score) and object-level AJI score metric.

### B. Deep Learning-based Semantic Segmentation Models

**Fully convolutional network (FCN)**: In [21], Long et al. proposed the FCN architecture that receives an input image with arbitrary size and produces pixel-wise predictions (i.e. segmentation mask), as shown in Fig. 3. They demonstrated that end-to-end and pixels-to-pixels deep convolutional networks could deliver promising results

with the semantic segmentation task. FCN includes deconvolutional layers to up-sample coarsely deep convolutional layer outputs to dense pixels of any desired resolution.

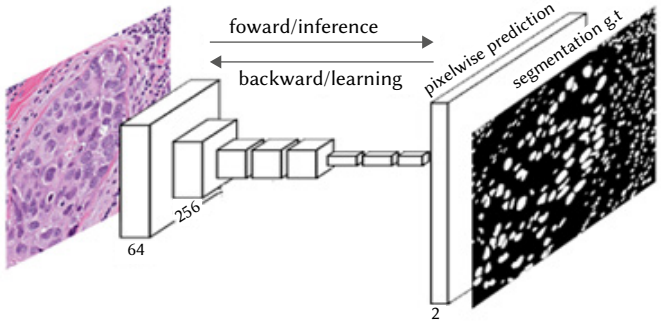



Fig. 3. Diagram of the FCN architecture.

The main idea behind FCN is to create a semantic segmentation network by adjusting state-of-the-art classification networks as such AlexNet [22], the VGG net [23], and GoogLeNet [24] into fully convolutional networks and transfer their representations to the segmentation task (e.g. use of fine-tuning techniques). It is worth noting that the structure of FCN allows generating segmentation maps for images of any resolution without employing fully connected layers, and therefore the FCN architecture is considered as one of the most innovative deep learning architecture that opened the door for several innovations in image segmentation based on deep learning [19]. Besides, skip connections are used to combine semantic information of different layers to produce accurate and detailed segmentation results. Specifically, skip connections enable information to flow, avoiding information loss because of other elements on deep learning architectures, such as max-pooling (down-sampling) and dropout layers. The common FCN architectures are FCN-32, FCN16, and FCN8 [21], which are based on VGG-16 backbone [23]. In our study, we use FCN8 as it gives the best performance.

**DenseNet**: Several subsequent approaches to semantic segmentation have been inspired by FCN [21], of them the architecture of FC-DenseNets [25]. In short, Jégou et al. [25] extended the architecture of densely connected convolutional networks (DenseNets) [26], which has achieved remarkable results on image classification tasks, to the semantic segmentation problem. In DenseNet, each layer is connected to other layers in a feed-forward fashion to facilitate the training process. Also, a feature reuse approach is implemented to enable all layers to access their preceding layers.

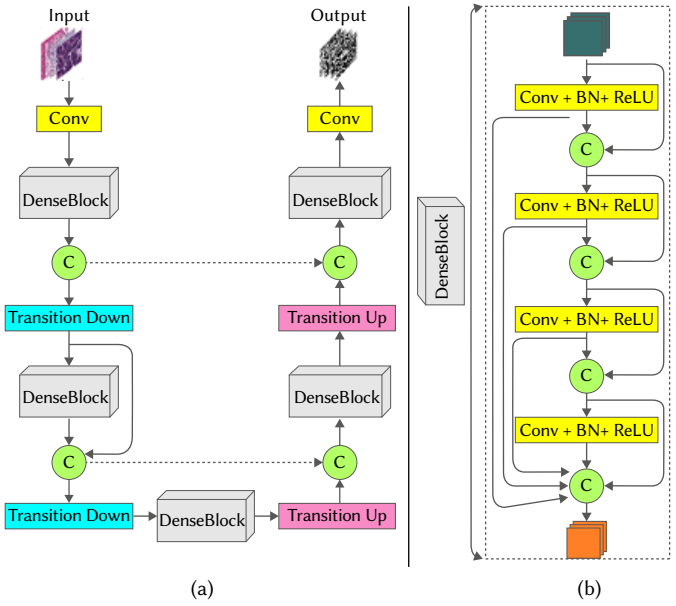



Fig. 4. Diagram of the FC-DenseNets architecture: (a) the two paths of FC-DenseNets architecture for semantic segmentation, and (b) a dense block with 4 layers.

Fig. 4 (a) shows the architecture of FC-DenseNet, which consists of the down-sampling path described in [26] and the upsampling path which allows recovering the full resolution of input images. As shown, in the down-sampling path the input to a dense block is concatenated with its output, which yields a linear growth in the number of feature maps. Notably, in the down-sampling path, the increase in the number of features is recompensed by decreasing in spatial resolution of each feature map after the pooling operation.

Fig. 4 (b) presents the architecture of the dense block. Of note, in FC-DenseNets an up-sampling process referred to as transition up. Transition up modules consist of a transposed convolution that up-samples the previous feature maps. The up-sampled feature maps are then concatenated to the ones coming by skip connection to form the input of a new dense block.

**U-Net**: Ronneberger et al. [28] proposed the U-Net architecture, which also is inspired by FCN architecture. The core idea of U-Net and its training strategy is based on the use of data augmentation methods to effectively learn from the available annotated samples. As shown in Fig. 5, the U-Net architecture is built based on the scheme of encoder-decoder networks, which enable capturing the contextual features from input images. The down-sampling path of U-Net (encoder network) follows the typical architecture of FCN to extract features. At each down-sampling step, the number of feature channels is doubled. The up-sampling path (decoder network) consists of deconvolution layers. Feature maps from the encoder network are concatenated with the corresponding ones of the decoder network to avoid losing pattern information (spatial information). Finally, a 1x1 convolution layer is used to generate the segmentation mask, where each pixel of the input channel is assigned to one of the classes.

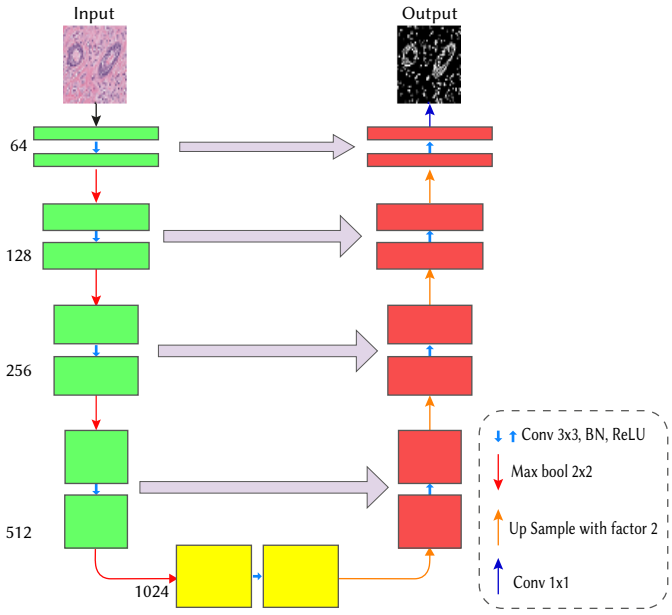



Fig. 5. Diagram of the U-Net architecture.

**SegNet**: In [29], Badrinarayanan et al. presented the SegNet architecture, which consists of an encoder network and a decoder network followed by a pixel-wise classification layer, as shown in Fig. 6. The encoder network of SegNet is similar to the first 13 convolutional

layers of VGG16 [23] without fully connected layers. The decoder network of SegNet comprises a hierarchy of decoders, where each one corresponds to an encoder layer. The decoder layers use max-pooling indices received from the corresponding encoder layers to perform non-linear up-sampling of the feature maps, which eliminates the need for learning to up-sample. The up-sampled maps are then convolved with trainable filters to produce dense feature maps.

Although U-Net and SegNet have similar architecture, U-Net does not reuse pooling indices, and it transfers the entire feature map to the corresponding decoder layers and concatenates them with the up-sampled feature maps, which costs more memory.

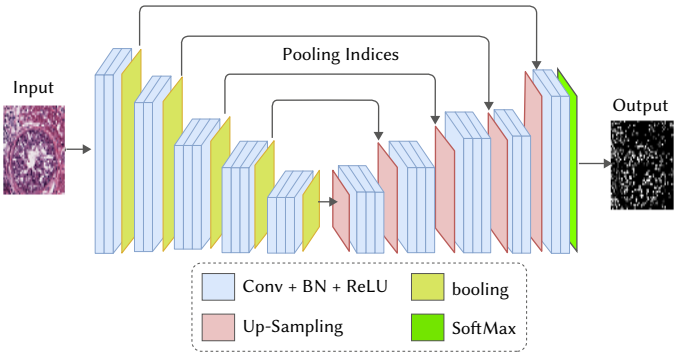



Fig. 6. Diagram of the SegNet architecture.

**Self-correction mechanism with CE2P network**: In [30], Li et al. introduced a training strategy called self-correction for human parsing (SCHP), which can iteratively improve the reliability of supervised labels as well as the learned models during the training process. The architecture used in SCHP is inspired by the CE2P architecture used in [31]-[32]. It consists of three main branches, namely parsing, edge, and fusion. The training strategy can be divided into two procedures: model aggregation, and label refinement (self-correction mechanism). A cyclically learning scheduler is used to produce reliable pseudo masks. Self-correction is performed iteratively by aggregating the current learned model with the former optimal one in an online manner.

In this study, a cyclically learning scheduler with warm restarts is used. In each cycle of the self-correction mechanism, we compute a set of weights (models), $W = \{\hat{w}_0, \hat{w}_1, \dots, \hat{w}_M\}$ and the corresponding predicted labels, $Y = \{\hat{y}_0, \hat{y}_1, \dots, \hat{y}_M\}$. After each training cycle, the current model weights $\hat{w}$ are combined with the weights of the previous cycle $\hat{w}_{n-1}$ to obtain new weights $\hat{w}_n$, as follows:

$$\hat{w}_n = \frac{n}{n+1}\hat{w}_{n-1} + \frac{1}{n+1}\hat{w} \tag{1}$$

Likewise, the predicted labels of the current cycle are combined with the labels of the previous cycle, as follows:

$$\hat{y}_n = \frac{n}{n+1}\hat{y}_{n-1} + \frac{1}{n+1}\hat{y} \tag{2}$$

where n refers to the current cycle number ($0 \le m \le M$) and $\hat{y}$ is the generated pseudo-labels (pseudo masks) with the model $\hat{w}_n$.

**PSPSegNet (PSP with SegNet)**: In [33], Zhao et al. developed the pyramid scene parsing network (PSPNet), which considers the strength of the global context of the image to enhance the local level predictions. The authors of PSPNet claim that FCN based architectures do not employ a suitable strategy to utilize the context of the whole image. Thus, they proposed a pyramid pooling module (PPM) to incorporate global contextual information. For each input image, PSPNet utilizes a pre-trained ResNet (feature extractor) to get feature maps. The feature maps from the feature extractor are pooled at four different scales corresponding to four different pyramid levels. Then, PPM is used to produce various sub-region representations, succeeded by up-sampling and concatenation layers to produce the final feature maps (i.e. the final representation) that comprise global and local contextual-information. The final representation is inputted into a convolution layer to produce the per-pixel prediction (i.e. the segmentation mask).

To improve the nuclei segmentation results, in this study, we present the PSPSegNet by combining PPM (the key component of PSPNet) with the SegNet architecture, as shown in Fig. 7. We use ResNet [34] as the encoder of SegNet, succeeded by the PPM module. The encoder feature maps are concatenated with the up-sampled outputs of the pyramid levels and then fed into the decoder of the SegNet to produce the segmented image.

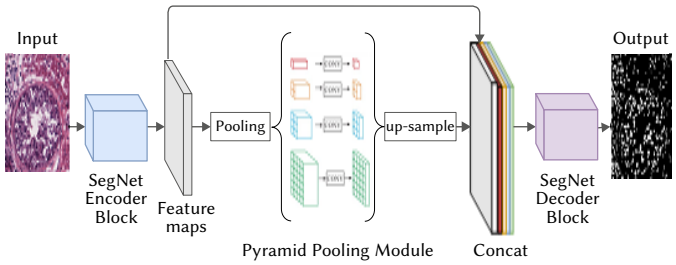



Fig. 7. Diagram of the PSPSegNet architecture.

In this study, we analyze the performance of the five models explained above (FCN, FC-DenseNet, U-Net, Self-correction, and PSPSegNet) with the nuclei segmentation task. We train them using the MoNuSeg dataset [27]. In the test phase, we separately assess the performance of each trained model.

### C. Model Evaluation

In this study, we use the aggregated Jaccard index (AJI) proposed in [27] and F1-score (dice score) to assess the performance of the nuclei segmentation methods. AJI is an extended version of the Jaccard index which divides the aggregated intersection cardinality by the aggregated union cardinality between the ground truth (G) and segmented masks. If AJI equals 1, it means that we obtain perfect nuclei segmentation results. AJI can be expressed, as follows:

$$AJI = \frac{\sum_{i=1}^{L} |G_i \cap P_j^*(i)|}{\sum_{i=1}^{K} |G_i \cup P_j^*(i)| + \sum_{K \in Ind} |P_K|} \tag{3}$$

where $G = U_{i=1,2,..K}\ G_i$ is the ground truth masks, $P = U_{i=1,2,..L}\ P_i$ are the prediction nuclei segmentation outputs, $P_j^*(i)$ is the connected component from the prediction output that maximizes the Jaccard index. *Ind* is the list of indices of pixels that do not belong to any element in ground truth (*G*).

The F1-score is the harmonic mean between the precision and recall. The F1-score is identical to the dice coefficient, which can be formulated, as follows:

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{4}$$

where TP, FP, FN refers to the true positive, false positive, and false-negative rates, respectively.

## IV. Experimental Results and Discussion

### A. Dataset

In our experiments, we use the MoNuSeg dataset [27], which contains 30 WSI images with annotations. MoNuSeg is a very

challenging dataset as it has WSI images of 7 organs (breast, kidney, colon, stomach, prostate, liver, and bladder) collected at different medical centers (i.e., various stains). The size of each WSI image is 1000×1000. Of the 30 WSI images, we use 23 WSI images for training the models and the rest for testing. Mainly, we keep one WSI image for each organ in the testing set.

### B. Implementation Details

As introduced in Section III.A, the computation cost of training deep learning models makes it very difficult to train them with the very high resolution of the input image (1000×1000 pixels). Therefore, in the preprocessing step, we apply the sparse stain normalization method of [20] on each WSI image and then rescale it to 1024×1024, and then we divide it into four non-overlapping sub-images of size 512×512. We augment the training data by cropping 200 512×512 patches from each WSI image randomly. Thus, the total training dataset has 4692 of 512×512 patches, and the testing set has 28 sub-images (7 WSI images × 4 splits). A GTX1080 with an 8GB memory GPU is used to run the experiments. All models are trained for 100 epochs, the stochastic gradient descent (SGD) is used as an optimizer

TABLE I. Summarization of Models Architecture

| FCN | FC-DenseNet | U-Net |
|---|---|---|
| **Input Layer, in_ch= 3** | | |
| **Feature Extraction Layers**<br>3x3 Conv+ReLU (2 layers) + MaxPool (S= 2), F= 64<br>3x3 Conv+ReLU (2 layers) + MaxPool (S= 2), F= 128<br>3x3 Conv+ReLU (3 layers) + MaxPool (S= 2), F= 256<br>3x3 Conv+ReLU (6 layers) + MaxPool (S= 2), F= 512 | **First Layer**<br>3x3 Conv, F= 48<br>**Down Sampling Layers**<br>DB (4 layers) + TD, F= 112<br>DB (5 layers) + TD, F= 192<br>DB (7 layers) + TD, F= 304<br>DB (10 layers) + TD, F= 464<br>DB (12 layers) + TD, F= 656 | **Contracting Path Layers**<br>3x3 Conv+BN+ReLU (2 layers) + MaxPool (S= 2), F=64<br>3x3 Conv+BN+ReLU (2 layers) + MaxPool (S= 2), F=128<br>3x3 Conv+BN+ReLU (2 layers) + MaxPool (S= 2), F=256<br>3x3 Conv+BN+ReLU (2 layers) + MaxPool (S= 2), F=512 |
| **Up-sample Layers**<br>3x3 DeConv + BN + ReLU, F= 512<br>3x3 DeConv + BN + ReLU, F= 256<br>3x3 DeConv + BN + ReLU, F= 128<br>3x3 DeConv + BN + ReLU, F= 64<br>3x3 DeConv + BN + ReLU, F= 32 | **Bottleneck Layers**<br>DB (15 layers), F= 896 | **Middle Layers**<br>3x3 Conv+BN+ReLU (2 layers) + MaxPool (S= 2), F=1024 |
| **Output (Classifier)_Layer**<br>1x1 Conv, C =2 | **Down Sampling Layers**<br>TU + DB (12 layers), F= 1088<br>TU + DB (10 layers), F= 816<br>TU + DB (7 layers), F= 578<br>TU + DB (5 layers), F= 284<br>TU + DB (4 layers), F= 256 | **Expanding Path Layers**<br>Upsample (scale=2) + 3x3 Conv+BN+ReLU (2 layers), F= 512<br>Upsample (scale=2) + 3x3 Conv+BN+ReLU (2 layers), F= 256<br>Upsample (scale=2) + 3x3 Conv+BN+ReLU (2 layers), F= 128<br>Upsample (scale=2) + 3x3 Conv+BN+ReLU (2 layers), F= 64 |
| | **Output Layer**<br>1x1 Conv, C = 2<br>SoftMax | **Output Layer**<br>1x1 Conv, C = 2 |

| PSPSegNet | SelfCorrection |
|---|---|
| **Input Layer, in_ch= 3** | |
| **First Layer**<br>7x7 Conv+BN+ReLU (1 layer) + MaxPool (S= 2), F=64 | **First Layers**<br>3x3 Conv+BN+ReLU (2 layers), F =64<br>3x3 Conv+BN+ReLU+ MaxPool (S =2), F=128 |
| **Encoders Layers (ResNet Backbone)**<br>3 x [(1x1 Conv+BN) + (3x3 Conv+BN)+(1x1 Conv+BN) + ReLU], F= 256<br>4 x [(1x1 Conv+BN) + (3x3 Conv+BN)+(1x1 Conv+BN) + ReLU], F= 512<br>23 x [(1x1 Conv+BN) + (3x3 Conv+BN)+(1x1 Conv+BN) + ReLU], F= 1024<br>3 x [(1x1 Conv+BN) + (3x3 Conv+BN)+(1x1 Conv+BN) + ReLU], F= 2048 | **Feature Extraction Layers (ResNet Backbone)**<br>3 x [(1x1 Conv+BN) + (3x3 Conv+BN)+(1x1 Conv+BN) + ReLU], F= 256<br>4 x [(1x1 Conv+BN) + (3x3 Conv+BN)+(1x1 Conv+BN) + ReLU], F= 512<br>23 x [(1x1 Conv+BN) + (3x3 Conv+BN)+(1x1 Conv+BN) + ReLU], F= 1024<br>3 x [(1x1 Conv+BN) + (3x3 Conv+BN)+(1x1 Conv+BN) + ReLU], F= 2048<br>**Context_Encoding:**<br>[AvgPool + (1x1 Conv+BN+ReLU)] (4 layers), F=4069<br>3x3 Conv+BN+ReLU, F= F= 2048 |
| **PPM Layer:**<br>[AvgPool + (1x1 Conv+BN+ReLU)] (4 layers), F=2048<br>1x1 Conv, F= 2048 | **Parsing Module_Layers**<br>1x1 Conv+BN+ReLU (4 layers), F =256<br>1x1 Conv, F= 2 |
| **Up-sample Layers**<br>5x5 Conv+BN (4 layers), F=1024, 512, 64, 32 respectively | **Edge_Module Layers**<br>1x1 Conv+BN (4 layers), F= 256<br>3x3 Conv, F= 2 |
| **Decoder Layers:**<br>1x1 Conv+BN+ReLU, F = 1024<br>1x1 Conv+BN+ReLU, F = 512<br>1x1 Conv+BN+ReLU, F = 64<br>1x1 Conv+BN+ReLU, F = 32 | **Output Fusion Module Layers:**<br>1x1 Conv+BN+ReLU, F=2<br>1x1 Conv, C= 2 |
| **Output_Layers:**<br>1x1 Conv+BN+ReLU, F=2<br>3x3 Conv, C= 2 | |

with an initial learning rate of 1e-1, a momentum of 0.99, and a weight decay of $1e^{-8}$. A batch size of two images is used.

Table I presents the architecture summarization of FCN, FC-DenseNet, U-Net, PSPSegNet, and Self-Correction models. Where *S*, *F* and *C* stand for stride, the number of feature maps (filters), and the number of output classes, respectively. With the FCN model, we follow the FCN8 architecture presented in [21] and use VGG-16 [23] as a feature extractor with its 13 convolutional layers. We decapitate VGG-16 by discarding the final classifier layer. An up-sampling with 5 deconvolution layers is employed after extracting the features. Finally, we utilize a 1x1 convolution with channel dimension 2 to predict scores for each nuclei class.

In the FC-Dense model, *DB* refers to the Dense-Block shown in Fig. 4. (b), where each layer in the block consists of a 3×3 convolution layer with a batch normalization layer followed by ReLU activation. *TD* refers to transition down operation using a 2×2 max-pooling layer with a stride of 2, and *TU* refers to the transition-up process using a 3x3 transpose convolution layer with a stride of 2.

We follow the same implementation of FC-DenseNet103 architecture presented in [25]. This architecture is built from 103 convolutional layers; the first convolutional layer is applied onto the input, 38 convolutional layers in the down-sampling path, 15 convolutional layers in the bottleneck, and 38 convolutional layers in the up-sampling path. Besides, 5 TD are used, each one containing a convolution, and 5 TU, each one containing a transposed convolution. Finally, a 1×1 convolution layer and a Softmax non-linearity are used to provide the per class distribution at each pixel.

In the U-Net model, two 3×3 convolutional layers followed by ReLU and 2×2 max-pooling operation with stride two are used in each encoder block. In each decoder block, two deconvolution layers are used, which are then concatenated with the corresponding feature maps of the encoder layers. The final layer of the decoder has a 1x1 convolution to map each 64-feature vector to two classes.

In the case of the PSPSegNet model, we use ResNet-101 architecture in the encoder. We implement a PPM between the encoder and the decoder. A bilinear up-sampling operation that consists of four 5×5 convolution layers with batch normalization is applied after PPM. The decoder includes four layers, where each consists of 1×1 convolution and batch normalization layers. Finally, a 1×1 convolution layer followed by a 3×3 convolution layer is used to provide the per class distribution at each pixel.

In the Self-Correction model, we use ResNet-101 [34] as a backbone of the feature extractor and use an ImageNet [35] pre-trained weights to commit with the same implementation in [30]. We adopt the PSP network [33] as a context encoding module. The parsing module and edge module comprise four 1×1 convolution layers with batch normalization and ReLU followed by one 1×1 convolution layer for the parsing module and one 3×3 convolution layer for the edge module. Finally, a fusion module is employed, which includes a 1×1 convolution layer with batch normalization and ReLU followed by a 1×1 convolution layer to predict scores for each nuclei class. Table II presents the architecture and the source code links of each model.

TABLE II. Models Architecture Backbones

| Model | Backbone | GitHub links |
|---|---|---|
| FCN | VGG-16 | *https://github.com/pochih/FCN-pytorch* |
| FC-Dense Net | - | *https://github.com/bfortuner/pytorch_tiramisu* |
| U-Net | - | *https://github.com/LeeJunHyun/Image_Segmentation* |
| PSPSegNet | ResNet-101 | *https://github.com/alexgkendall/SegNet-Tutorial* |
| Self-Correction | ResNet-101 | *https://github.com/PeikeLi/Self-Correction-Human-Parsing* |

## C. Results and Discussion

Table III shows the segmentation results of the nuclei segmentation models with the MoNuSeg dataset in terms of the dice coefficient (F1-score) and AJI score. As shown, FCN obtains F1-score of 0.8467 and AJI of 0.6418, which are lower than the other four models. FC-Dense Net and U-Net achieve improvements on the F1-score of 1.2% and 1.56%, respectively, when compared to FCN. Besides, they give gains on the AJI of 2.28% and 3.38%, respectively. PSPSegNet obtains the best results with 3.48% improvement on the F1-score and 6.62% improvement on AJI, thanks to PPM that encourages the PSPSegNet model to learn global context features of WSI images. The F1-score and AJI of PSPSegNet are 0.26% and 0.66% higher than the ones of Self-Correction. This analysis reveals that both PSPSegNet and Self-Correction could be used to get suitable nuclei segmentation results.

TABLE III. Comparison Between the Nuclei Segmentation Models

| | F1-Score | AJI |
|---|---|---|
| **FCN** | 0.8467 | 0.6418 |
| **FC-Dense Net** | 0.8587 | 0.6646 |
| **U-Net** | 0.8623 | 0.6756 |
| **PSPSegNet** | **0.8815** | **0.7080** |
| **Self-Correction** | 0.8792 | 0.7014 |

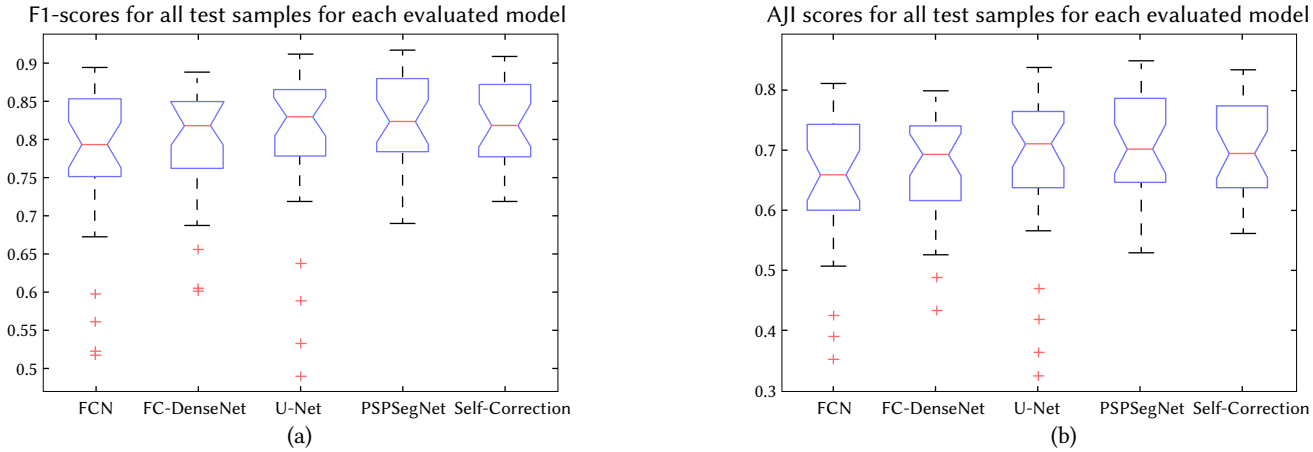
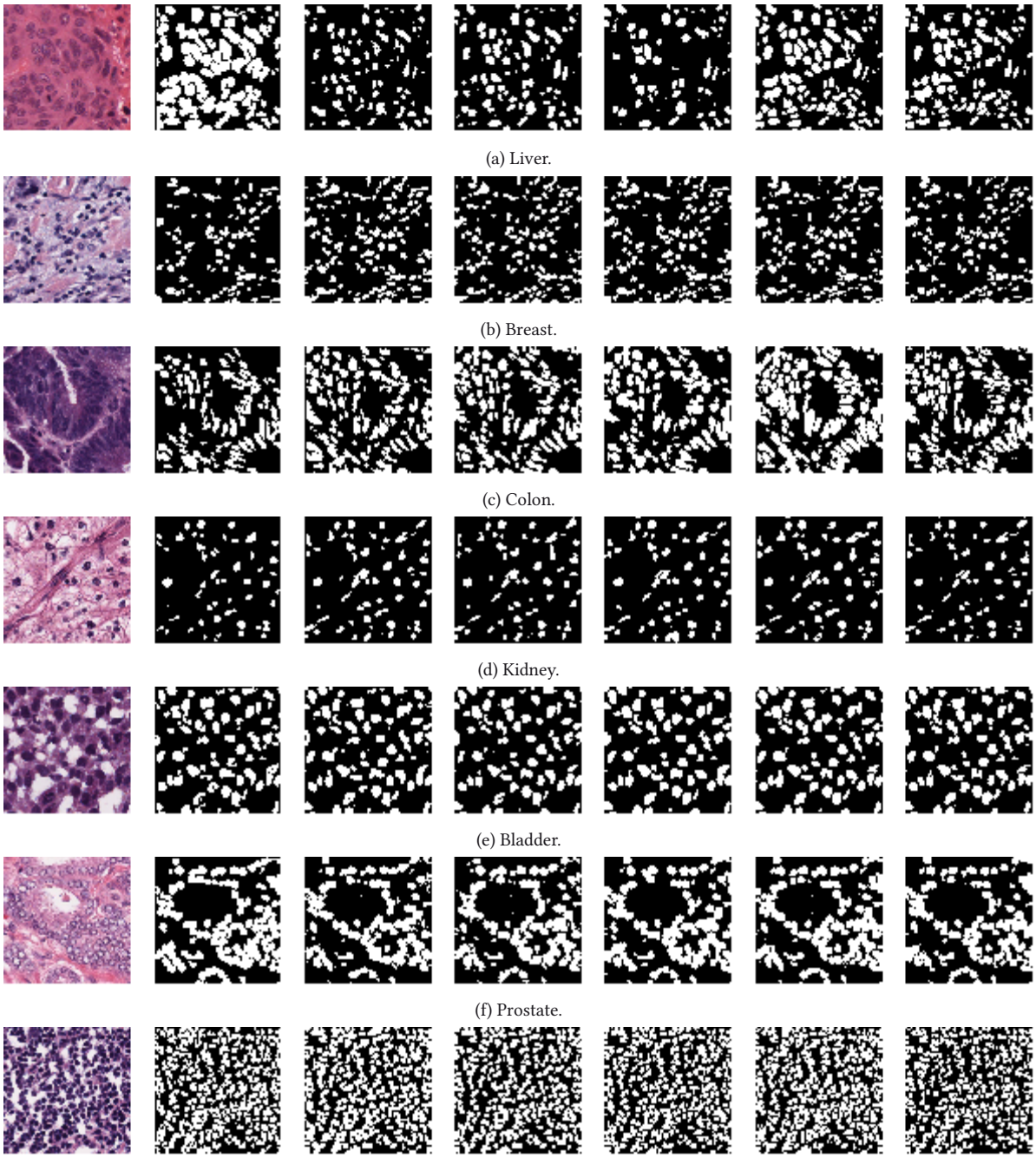



Fig. 8. Boxplots of F1-score and AJI of the five nuclei segmentation models: (a) F1-score, and (b) AJI.

Fig. 8 shows the boxplots of F1-score and AJI for all nuclei segmentation models. Given the scores of test images with a particular model, a boxplot can be displayed based on a five-number summary: the minimum, the maximum, the sample median, and the first and third quartiles. In Fig. 8, the red horizontal line refers to the sample median. As shown, the PSPSegNet and Self-Correction models have not any outliers on F1-score and AJI values. U-Net model has the highest median F1-score and AJI, but it produces the highest number of outliers on both evaluation metrics. However, FCN has AJI less than U-Net; it has a lower number of outliers. As we can see, PSPSegNet and Self-Correction models almost achieve the same median AJI and F1-score values. PSPSegNet achieves the maximum AJI and F1-scores when compared to other models, while FCN produces the minimum values.

Fig. 9 presents samples of segmented WSI images of different organs. We can notice that segmentation results can vary from one organ to another. For example, the WSI image of the liver organ (Fig. 9 (a)) has several big nuclei and some of them are overlapped. As shown in Fig. 9 (a) Col. 6, the PSPSegNet model accurately segments the cell nuclei with an AJI score 0.635, while the FCN model gives the worst segmentation results with an AJI score of 0.349 (Fig. 9 (a), Col. 3). We believe that the good performance of PSPSegNet is a result of the employment of PPM that integrates multi-scale maps in the middle of the model to learn WSI image context features. The same conclusion can be said for the colon organ WSI image (Fig. 9 (c)) and the bladder WSI image (Fig. 9 (e)).



(a) Liver.

(b) Breast.

(c) Colon.

(d) Kidney.

(e) Bladder.

(f) Prostate.

(g) Stomach.

Fig. 9. Segmentation results of all models with different organs (liver, breast, colon, kidney, prostate, bladder, and stomach). The first and second columns represent the original input image and ground truth mask, respectively. Columns 3-7 represent the output mask of FCN, FC-DenseNet, U-Net, PSPSegNet, and SelfCorrection, respectively.

Fig. 9 (b) shows a WSI image of the breast organ that has a noticeable color stain variation with a lot of overlapped nuclei. As shown in Fig. 9 (b) Col. 7, the Self-Correction model obtains the best performance with an AJI score of 0.655, thanks to the cyclically learning scheduler that enhances the segmentation results. The same conclusion can be said for the WSI image of the kidney organ shown in Fig. 9 (d). Fig. 9 (g) presents a WSI image of the stomach organ that has a dense number of nuclei. As we can see, all models produce good segmentation results. Specifically, the PSPSegNet model achieves the best results with an AJI score of 0.829. It is worth noting that all models obtain good segmentation results with stomach, liver and breast WSI images. However, in the case of the most complex WSI images that have color stain variation and overlapped nuclei, PSPSegNet and Self-Correction models produce the best segmentation results.

As nuclei segmentation is crucial to cell counting, it is interesting to study the performance of the five segmentation models with this task. In this regard, Fig. 10 presents a comparison between the number of cell nuclei in the predicted masks of each segmentation model and the ground-truth. To count the number of cell nuclei in each mask, we employ the connected component algorithm. In this experiment, we empirically set a threshold of 70 pixels for the minimum area of cells. As shown in Fig. 10, the number of cell nuclei obtained by the PSPSegNet model is a bit higher than the ones of the ground-truth, while the number of cell nuclei obtained by the Self-Correction model is close to the ground-truth. In turn, FCN, FC-Dense, and U-Net models have a lower number of cells than the ground-truth.

Fig. 11 shows a comparison between the number of parameters of FCN, FC-DenseNet, U-Net, PSPSegNet, and self-correction models. As shown, there is a noticeable variation in complexity between the nuclei segmentation models. It is worth noting that the good results achieved with the PSPSegNet model (F1-score 0.882 and AJI score 0.708) cost a massive number of parameters that exceed 122 million.

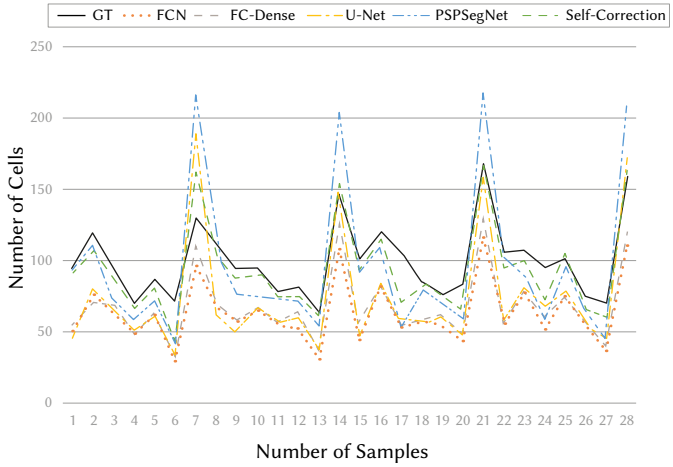



Fig. 10. Comparison between the number of cell nuclei in the predicted masks of each segmentation model and the ground-truth.

The Self-Correction model has 66 million parameters, which is almost half of PSPSegNet while achieving acceptable segmentation results (F1-score 0.879 and AJI score 0.701). However, the number of parameters of the FC-DenseNet model is less than the FCN model (around 9 million); it achieves a better F1-score and AJI. The FC-DenseNet model is also better than the U-Net model in terms of the number of parameters without a big difference in the segmentation results.

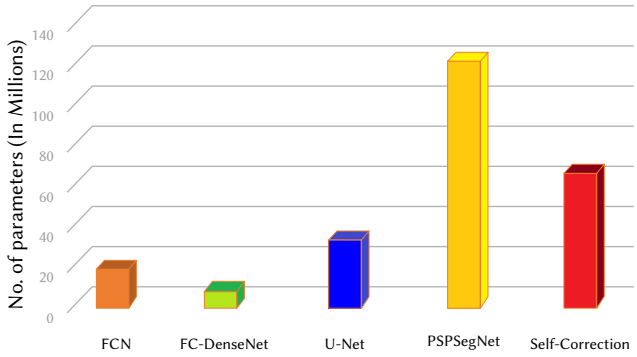



Fig. 11. Comparison between the number of parameters of FCN, FC-DenseNet, U-Net, PSPSegNet, and self-correction models.

In this study, we also ensemble the prediction masks of the top three models (U-Net, PSPSegNet, and self-Correction) using a simple pixel-wise aggregation function, as follows:

$$Ens = OR\,[U_{mask}, AND\,(PSP_{mask}, S_{mask})] \quad (5)$$

where $U_{mask}$, $PSP_{mask}$, and $S_{mask}$ are the prediction masks of U-Net, PSPSegNet, and Self-Correction models, respectively. *Ens* is the ensemble output. This ensemble method increases the AJI score to 0.7103 (0.23% improvement on the AJI compared to PSPSegNet).

Indeed, pathologists prefer to use easy and friendly software to segment cells from histopathology images. One of the most popular software is ImageFIJI [3]. Fig. 12 shows the segmentation results of the ImageFIJI program. As shown, the segmentation result of ImageFIJI is worse than the one of the proposed ensemble model (U-Net, PSPSegNet, and Self-Correction models). Also, ImageFIJI gives an AJI score of 0.533, which is much lower than the one of FCN model (the worst nuclei segmentation model presented in this study). Therefore, PSPSegNet could be a proposing nuclei segmentation method for pathologists.

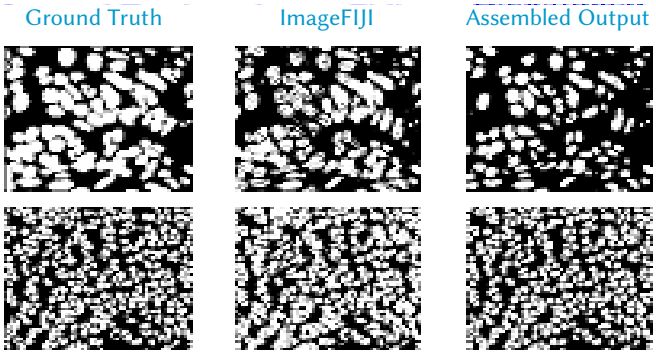



Fig. 12. Segmentation results of (middle) ImageFIJI software, and (right) ensemble of U-Net, PSPSegNet, and Self-Correction models.

In the literature, several studies have also used the MoNuSeg dataset and evaluated the segmentation models using F1-score and AJI. For instance, the study of [12] proposed a deep conditional generative adversarial network (cGAN) model, the study of [11] proposed a weakly supervised deep nuclei segmentation model, and the study of [7] presented the bending loss regularized network, to segment nuclei in WSI images. Mahmood et al. [12] obtained an F1 score of 0.866 and an AJI score 0.721, Qu et al. [11] achieved an F1 score of 0.778 and an AJI score of 0.505, and Wang et al. [7] obtained an AJI score 0.641. Based on the results mentioned above, we can conclude that PSPSegNet and self-correction approach outperforms all these approaches in term of pixel-level F1-score. Besides, the results achieved in [11] are much lower than the ones of PSPSegNet, and Self-

Correction models in terms of F1-score and AJI, noting that weakly supervised nuclei segmentation using points annotation technique may not be appropriate for cell nuclei segmentation. In terms of object-level AJI score, we can see that the results of [12] exceed our models, noting that this approach depends on synthetic training data that may not produce very accurate cell shapes.

To end, the models presented in this study can overcome the challenges that existing nuclei methods face. Specifically, PSPSegNet achieves promising performance in terms of the F1-score AJI scores. Thus, PSPSegNet could be a feasible nuclei segmentation tool for pathologists. It is worth noting that the segmentation model presented in this paper can be used to segment region of interest in several medical image modalities, such as the segmentation of nipples in thermograms [35], segmentation of pectoral muscle in mammograms [36], and vessel segmentation in fundus images [37].

## V. Conclusion

In this paper, we have presented promising deep semantic nuclei segmentation models in WSI images of different organs and collected from various clinics. To overcome the challenges that existing nuclei segmentation models face, we have sought the efficacy of pertinent deep learning models with nuclei segmentation task. Besides, we have consolidated robust deep learning architectures to build an efficient deep learning nuclei segmentation model (named PSPSegNet).

To demonstrate the performance of the nuclei segmentation models, we have used a well-known multi-organ WSI image dataset that includes WSI collected from different organs and scanners. We have comprehensively compared the performance of the nuclei segmentation models and exiting software in terms of the F1-score metric, object-level AJI metric, and the number of trainable parameters. The proposed PSPSegNet model achieved the highest performance with a pixel-level F1-score of 0.8815 and an object-level AJI score of 0.7080. PSPSegNet achieves promising results, but it has 122 million trainable parameters. In comparison with FCN, PSPSegNet achieved 3.48% improvement on the F1-score and 6.62% improvement on AJI.

Interesting results have been obtained with the Self-correction model with an F1-score of 0.8815 and AJI-score of 0.7080 with almost 67 million trainable parameters. Of note, the number of trained parameters of FCN, FC-DenseNet, and U-Net models ranges from 9 to 34 million, but they obtained lower segmentation performance than PSPSegNet with F1-scores of 0.847, 0.859, 0.862 respectively, and AJI scores of 0.641, 0.665, 0.676 respectively. Also, we have compared the performance of PSPSegNet with existing software (ImageFIJI), noting that PSPSegNet achieves better results. The experimental results emphasize that the PSPSegNet model could be used in the cell counting task.

The future work will include several extensions of the current study:

- The use of different aggregation strategies to combine the individual nuclei segmentation models.
- Incorporation of stain normalization techniques into the deep learning framework with different strategies.
- Converting the WSI images to other coordinate systems, such as the log-polar coordinates [39] and the curvilinear coordinates [37] to improve the nuclei segmentation results.

## Acknowledgment

M. Abdel-Nasser and D. Puig are partially supported by the Spanish Government through Project under Grant PID2019-105789RB-I00.

## References

[1] Moen, E., Bannon, D., Kudo, T. et al. "Deep learning for cellular image analysis," i*n Nat Methods*, vol. 16, no. 12, pages. 1233–1246, 2019, doi: 10.1038/s41592-019-0403-1.

[2] Carpenter, A.E., Jones, T.R., Lamprecht, M.R. et al. "CellProfiler: image analysis software for identifying and quantifying cell phenotypes," *in Genome Biol*, vol 7, R100, 2006, doi: 10.1186/gb-2006-7-10-r100.

[3] Schindelin, J., Arganda-Carreras, I., Frise, E. et al., "Fiji: an open-source platform for biological-image analysis," *in Nat Methods*, vol. 9, no. 7, pages. 676- 682, 2012.

[4] Niazi, Muhammad Khalid Khan et al. "Digital pathology and artificial intelligence," *in The Lancet. Oncology*, vol. 20, no. 5, pages e253-e261, 2019, doi: 10.1016/S1470-2045(19)30154-8

[5] P. Naylor, M. Laé, F. Reyal and T. Walter, "Nuclei segmentation in histopathology images using deep neural networks," *in 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, Melbourne, VIC, 2017, pp. 933-936, doi: 10.1109/ISBI.2017.7950669.

[6] P. Naylor, M. Laé, F. Reyal and T. Walter, "Segmentation of Nuclei in Histopathology Images by Deep Regression of the Distance Map," *in 2019 IEEE Transactions on Medical Imaging*, vol. 38, no. 2, pp. 448-459, Feb. 2019, doi: 10.1109/TMI.2018.2865709.

[7] H. Wang, M. Xian and A. Vakanski, "Bending Loss Regularized Network for Nuclei Segmentation in Histopathology Images," *in 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, Iowa City, IA, USA, 2020, pp. 1-5, doi: 10.1109/ISBI45749.2020.9098611.

[8] Graham, Simon Matthew, Quoc Dang Vu, Shan-e-Ahmed Raza, Jin Tae Kwak and Nasir M. Rajpoot. "XY Network for Nuclear Segmentation in Multi-Tissue Histology Images," in ArXiv abs/1812.06499 2018.

[9] Al-Kofahi, Y., Zaltsman, A., Graves, R. et al. "A deep learning-based algorithm for 2-D cell segmentation in microscopy images," *in BMC Bioinformatics*, vol. 19, no. 365, 2018, doi: 10.1186/s12859-018-2375-z

[10] Cui, Y., Zhang, G., Liu, Z. et al. "A deep learning algorithm for one-step contour aware nuclei segmentation of histopathology images," *in Med Biol Eng Comput*, vol. 57, pages 2027–2043, 2019, doi: 10.1007/s11517-019-02008-8

[11] H. Qu et al., "Weakly Supervised Deep Nuclei Segmentation Using Partial Points Annotation in Histopathology Images," *in 2020 IEEE Transactions on Medical Imaging*, doi: 10.1109/TMI.2020.3002244.

[12] F. Mahmood et al., "Deep Adversarial Training for Multi-Organ Nuclei Segmentation in Histopathology Images," *in 2019 IEEE Transactions on Medical Imaging*, 2019, doi: 10.1109/TMI.2019.2927182.

[13] Zhou, Y., Onder, O.F., Dou, Q., Tsougenis, E., Chen, H., Heng, P.A., "CIA-Net: Robust Nuclei Instance Segmentation with Contour-Aware Information Aggregation," *International Conference on Information Processing in Medical Imaging*, 2019, pp. 682-693.

[14] Wang, E.K., Zhang, X., Pan, L., Cheng, C., Dimitrakopoulou-Strauss, A., Li, Y., Zhe, N., "Multi-Path Dilated Residual Network for Nuclei Segmentation and Detection," *Cells Journal, Multidisciplinary Digital Publishing Institute*, 2019, vol. 8, no. 5, page 499.

[15] D. S. Mercadier, B. Besbinar and P. Frossard, "Automatic Segmentation of Nuclei in Histopathology Images Using Encoding-decoding Convolutional Neural Networks," *in 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, 2019, pp. 1020-1024, doi: 10.1109/ICASSP.2019.8682502.

[16] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh and J. Liang, "UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation," in *2020 IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856-1867, June 2020, doi: 10.1109/TMI.2019.2959609.

[17] Jung, H., Lodhi, B., Kang, J., "An automatic nuclei segmentation method based on deep convolutional neural networks for histopathology images," *BMC Biomedical Engineering Journal*, vol.1, no. 24, 2019, doi: 10.1186/s42490-019-0026-8.

[18] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," *in 2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, 2017, pp. 2980-2988, doi: 10.1109/ICCV.2017.322.

[19] Lateef, Fahad, and Yassine Ruichek. "Survey on semantic segmentation using deep learning techniques," *in Neurocomputing*, vol. 338, pp. 321-348, 2019.

[20] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K.

Steiger, A. M. Schlitter, I. Esposito, and N. Navab, "Structurepreserving color normalization and sparse stain separation for histological images," *in 2016 IEEE transactions on medical imaging*, vol. 35, no. 8, pp. 1962–1971, 2016.

[21] J. Long, E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation," *in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 2015, pp. 3431-3440, doi: 10.1109/CVPR.2015.7298965.

[22] Krizhevsky, Alex & Sutskever, Ilya & Hinton, Geoffrey. "ImageNet Classification with Deep Convolutional Neural Networks," *Neural Information Processing Systems*, vol. 25, 2012.

[23] Simonyan, Karen and Zisserman, Andrew. "Very Deep Convolutional Networks for Large-Scale Image Recognition," *in Computing Research Repository (CoRR)*. [Online]. Available: http://arxiv.org/abs/1409.1556.

[24] C. Szegedy et al., "Going deeper with convolutions," *in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 2015, pp. 1-9, doi: 10.1109/CVPR.2015.7298594.

[25] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero and Y. Bengio, "The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation," *in 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI, 2017, pp. 1175-1183, doi: 10.1109/CVPRW.2017.156.

[26] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, "Densely Connected Convolutional Networks," *in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 2261-2269, doi: 10.1109/CVPR.2017.243.

[27] N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane and A. Sethi, "A Dataset and a Technique for Generalized Nuclear Segmentation for Computational Pathology," in 2017 *IEEE Transactions on Medical Imaging*, vol. 36, no. 7, pp. 1550-1560, July 2017, doi: 0.1109/TMI.2017.2677499.

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation", *International Conference on Medical image computing and computer-assisted intervention*. Springer, vol 9351, pp. 234–241, 2015, doi: 0.1007/978-3-319-24574-4_28.

[29] V. Badrinarayanan, A. Kendall and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *in 2017 IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481-2495, 1 Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.

[30] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia, "Pyramid Scene Parsing Network," *in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 6230-6239, doi: 10.1109/CVPR.2017.660.

[31] Li, P., Xu, Y., Wei, Y., and Yang, Y. "Self-correction for human parsing," 2019, arXiv:1910.09777.

[32] Ruan, T., Liu, T., Huang, Z., Wei, Y., Wei, S., and Zhao, Y., "Devil in the Details: Towards Accurate Single and Multiple Human Parsing". *In 2019 Proceedings of the AAAI Conference on Artificial Intelligence*, 2019 volume 33, pages 4814–4821.

[33] Abdel-Nasser, M., Saleh, A., and Puig, D. "Channel-wise Aggregation with Self-correction Mechanism for Multi-center Multi-Organ Nuclei Segmentation in Whole Slide Imaging". *In 2020 Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, VISIGRAPP 2020, Vol. 4: VISAPP, Valletta, Malta, February 27-29, 2020 (pp. 466–473).

[34] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[35] J. Deng, W. Dong, R. Socher, L. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," *in 2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, 2009, pp. 248-255.

[36] Abdel-Nasser, M., Saleh, A., Moreno, A., and Puig, D. "Automatic nipple detection in breast thermograms," *Expert Systems with Applications*, vol. 64, pp. 365-374, 2016, doi: 10.1016/j.eswa.2016.08.026

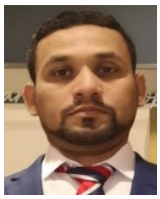
[37] Abdel-Nasser, M., Moreno, A., and Puig, D. "Temporal mammogram image registration using optimized curvilinear coordinates," *Computer methods and programs in biomedicine*, vol. 127, pp. 1-14, 2016. doi: 10.1016/j.cmpb.2016.01.019

[38] WANG, Xiaohong; JIANG, Xudong; REN, Jianfeng. "Blood vessel segmentation from fundus image by a cascade classification framework," *in Pattern Recognition*, 2019, vol. 88, p. 331-341.

[39] Roberts, N., Magee, D., Song, Y., Brabazon, K., Shires, M., et. al. "Toward routine use of 3D histopathology as a research tool," *The American journal of pathology*, vol. 180, no. 5, pp. 1835-1842, 2012.

Loay Hassan

He received the B.Sc. and M.Sc. degrees from Aswan University, in 2010 and 2015, respectively. Currently, he is a PhD student in Aswan university. He worked as an assistant researcher at Center for Artificial Intelligence and Robotics (CAIRO), Aswan University for 4 years. Currently, he is working as a part-time teacher assistant in Department of computer science at Arab Academy for Science, Technology & Maritime Transport, south valley branch.



Adel Saleh

He is currently a researcher at Gaist Solutions Ltd (UK). He received his Ph.D. in Computer Engineering from Universitat Rovira i Virgili (Spain) in 2019 and his master degree in computer science from VSTU (Russia). Dr. Saleh has published more than 20 papers in international journals and conferences. His research interests include images and videos analysis based on deep learning.

Mohamed Abdel-Nasser

He received the Ph.D. degree in Computer Engineering from the University Rovira i Virgili (URV) in 2016. He has been a postdoc researcher at the URV since 2018, and Assistant Professor at the Electrical Engineering Department, Aswan University (Egypt) since 2016. In 2017, he has received the Marc Esteva Vivanco prize for the best Ph.D. dissertation on Artificial Intelligence. He has participated in several projects funded by the European Union and the Government of Spain. He has published more than 65 papers in international journals and conferences. He was the manager of the E-learning & digital library center at Aswan University (2018). His research interests include the application of machine learning and deep learning to several real-world problems, including medical image analysis, smart road environment, smart grid analysis and time-series forecasting.

Osama A. Omer

He received the B.Sc. and M.Sc. degrees from South Valley University, in 2000 and 2004, respectively, and the Ph.D. degree from the Tokyo University of Agriculture and Technology, in 2009. He spent six months as a Postdoctoral Researcher with the Medical Engineering Department, Luebeck University, Germany. Also, he spent three months as a Postdoctoral Researcher with Kyushu University, Japan. He spent six months as an R&D Scientist Engineer with the NOKIA R&D Center, Tokyo, Japan, in 2008. He is currently a Professor with the Electrical Engineering Department, Aswan University. Prof Omer has published more than 70 papers in international journals and conferences. His research interests include medical imaging, super-resolution, image/video coding, and wireless communications.

Domenec Puig

He received the M.S. and Ph.D. degrees in computer science from the Polytechnic University of Catalonia, Barcelona, Spain, in 1992 and 2004, respectively. In 1992, he joined the Department of Computer Science and Mathematics, Rovira i Virgili University (URV), Tarragona, Spain, where he is currently working as a Professor since 2017. He has been the Head of the Intelligent Robotics and Computer Vision Group, Rovira i Virgili University since 2006. Prof. Puig is also the Vice-Rector of the URV. He is the principal investigator (PI) of several projects funded by the European Union and the Government of Spain. He has published more than 200 papers in international journals and conferences His research interests include artificial intelligence and mobile robotics.

# Application of Artificial Intelligence Algorithms Within the Medical Context for Non-Specialized Users: the CARTIER-IA Platform

Francisco José García-Peñalvo[1], Andrea Vázquez-Ingelmo[1], Alicia García-Holgado[1], Jesús Sampedro-Gómez[2], Antonio Sánchez-Puente[2], Víctor Vicente-Palacios[3], P. Ignacio Dorado-Díaz[2], Pedro L. Sánchez[2] *

[1] GRIAL Research Group, University of Salamanca (Spain)
[2] Cardiology Department, Hospital Universitario de Salamanca, SACyL. IBSAL, Facultad de Medicina, University of Salamanca, and CIBERCV (ISCiii) (Spain)
[3] Philips Healthcare (Spain)

Received 1 March 2021 | Accepted 19 April 2021 | Published 13 May 2021



## Abstract

The use of advanced algorithms and models such as Machine Learning, Deep Learning and other related approaches of Artificial Intelligence have grown in their use given their benefits in different contexts. One of these contexts is the medical domain, as these algorithms can support disease detection, image segmentation and other multiple tasks. However, it is necessary to organize and arrange the different data resources involved in these scenarios and tackle the heterogeneity of data sources. This work presents the CARTIER-IA platform: a platform for the management of medical data and imaging. The goal of this project focuses on providing a friendly and usable interface to organize structured data, to visualize and edit medical images, and to apply Artificial Intelligence algorithms on the stored resources. One of the challenges of the platform design is to ease these complex tasks in a way that non-AI-specialized users could benefit from the application of AI algorithms without further training. Two use cases of AI application within the platform are provided, as well as a heuristic evaluation to assess the usability of the first version of CARTIER-IA.

## Keywords

Information System, Medical Data Management, Medical Imaging Management, Artificial Intelligence, Health Platform.

DOI: 10.9781/ijimai.2021.05.005

## I. Introduction

ARTIFICIAL Intelligence (AI) algorithms have grown in popularity and increased its range of uses over the years. The possibility of applying them to different problems and contexts provide a wide support in complex scenarios in which data is continuously being generated.

One of these complex scenarios is the medical context. These algorithms and approaches are becoming very relevant when analyzing medical data [1]. However not only structured or tabular data can be involved in this context; medical imaging are also crucial resources within the medical domain.

The analysis of medical imaging involves complex tasks such as disease detection, segmentation, assessment of organ functions, etc. [2]-[4]. In this sense, artificial intelligence algorithms can provide support to these tasks with similar performance compared to human skills [5].

However, as introduced, data is being continuously generated in medical scenarios, which makes its management a convoluted responsibility. In fact, not only several data sources can be involved, but also different data structures. This data heterogeneity is a challenge both for its management and the application of AI algorithms.

Because of this, one of the main challenges of applying AI algorithms in real medical scenarios relies on the unification and accessibility of the generated data. For this reason, information systems are required to gather, clean, organize and structure data in order to apply AI algorithms in a friendly, secure and anonymized manner.

This work presents a platform for the management of structured data and imaging resources in the medical context with advanced features such as their visualization, edition and application of AI on the stored resources.

Powerful tools such as information dashboards can be easily integrated in the platform [6], [7] to explore structured data. This kind of tools provide support to knowledge generation, which is very relevant in this context [8].

On the other hand, DICOM editors and AI integration are also crucial components, which allow the modification and advanced exploration of imaging data.

* Corresponding author.
E-mail addresses: fgarcia@usal.es (F. J. García-Peñalvo), andreavazquez@usal.es (A. Vázquez-Ingelmo), aliciagh@usal.es (A. García-Holgado), jmsampedro@saludcastillayleon.es (J. Sampedro-Gómez), asanchezpu@saludcastillayleon.es (A. Sánchez-Puente), victor.vicente.palacios@philips.com (V. Vicente-Palacios), pidorado@saludcastillayleon.es (P. I. Dorado-Díaz), plsanchez@saludcastillayleon.es (P. L. Sánchez)

The starting point of this project stems from the need of using a collaborative platform to gather these heterogeneous types of data. Unifying medical data sources through a collaborative platform eases their exploration and analysis, as well as enabling the possibility of sharing knowledge across different projects.

In this case, the platform was built for research purposes in the field of cardiology, but its flexibility enables its use for other fields in which structured data and medical imaging need to be unified.

In fact, the features of this platform can also provide support to educational purposes, in which the application of AI scripts is guided and explained to novice or non-specialized users [9], [10].

Relying on a web-collaborative platform also allows the integration of artificial intelligence algorithms. In fact, cardiac imaging is particularly interesting for the application of AI algorithms, since many tasks are related to the assessment of volumes, distances and motion of different structures in the heart and, in this regard, deep learning techniques have been proven to achieve good results [11].

Cardiac imaging is usually composed of DICOM (Digital Imaging and Communication On Medicine) files [12] from echocardiographic, magnetic resonance, or computed tomography, among the most important. Other online medical imaging platforms implement the DICOM protocol and are available for these purposes, some of them even in an open-source format [13].

On the other hand, there exist solutions based on application programming interfaces with pretrained models' repositories for use in the medical imaging field. Nevertheless, these repositories are mostly oriented towards advanced users with expertise in programming and data science knowledge [14].

As can be seen, different solutions arise to manage medical imaging and execute AI algorithms over them. However, in this scenario, it is necessary not only to unify data sources, but also these kinds of services.

For these reasons, the development of a technological ecosystem [15], [16] is an appropriate solution to merge both functionalities into a user-friendly web-based interface.

The CARTIER-IA platform can be seen as a technological ecosystem that support all data-management related tasks (including structured data and medical imaging collection) and also enable both healthcare professionals and data scientists to apply AI models to the stored images.

Deep learning and machine learning models can be stored by AI developers through Python scripts, as it will be detailed in section 2.C. Using this approach, scripts can be executed through the web interface by any user interested in analyzing the image, with the goal of providing the benefits of these scripts without requiring python-programming skills nor advanced knowledge regarding Artificial Intelligence algorithms.

In this respect, given the fact that not every user is skilled in programming AI algorithms, the platform needs a user-centered approach to provide friendly interfaces and bring AI-driven tasks closer to non-specialized users.

In this work we present the integration of AI algorithms into the CARTIER-IA platform's DICOM viewer and editor. A heuristic evaluation of the tool is provided to test its usability and improve the image processing and AI application workflow with the goal of offering better user experience.

The structure of this paper is as follows: section II outlines the technical details of the platform as well as the heuristic evaluation methodology, section III explains the main functionality blocks of the CARTIER-IA platform, section IV describes two use cases of the AI integration within the platform and section V provides the heuristic evaluation results regarding the image editor and script application tool. Finally, section VI and section VII discuss the results and present the conclusions, respectively.

## II. Methodology

### A. Technical Details of the Platform

The platform relies on different technologies and frameworks which are integrated using a client-server architecture.

The front-end employs HTML, CSS, and JavaScript to send data to the server. On the other hand, the DICOM viewer and editor is also located at the front end, and it is implemented through the Cornerstone.js library.

On the other hand, the back end performs more complex tasks to fulfil the requirements of the platform, such as the data storage, data processing and an Artificial Intelligence environment.

The technology employed to implement this client-server approach as a web application is Django, a Python-based web framework [17]. The web application is also connected through web requests to other services such as a REDCap instance to manage additional projects and information.

Due to the necessity of pre-validate DICOM images and structured data, upload processes can be time-consuming tasks. For this reason, job queries have been implemented to carry out these data uploads asynchronously as background jobs. This allow users to navigate the platform while their data is uploading.

Finally, to implement the integrated AI environment, the back end is supported by libraries such as OpenCV and TensorFlow, in order to enable the execution of deep learning models and other AI-related scripts.

### B. Usability Study: Heuristic Evaluation

Integrating complex tasks such as AI algorithms in a web interface in which other diverse functionalities are involved is a challenge, especially regarding providing a good user experience.

For this reason, the platform needs to be thoroughly tested in terms of its usability. One of the preliminary studies that has been carried out to identify interface design weaknesses in the platform's first version is a heuristic evaluation.

Although there are several heuristics sets to perform heuristic evaluations, the most popular are the ten heuristics by Nielsen [18]. There are also specific heuristics related to the medical domain, but there aren't focused on this kind of platforms (they are mostly related to the evaluation of Electronic Health Records [19], [20]).

However, due to the fact that CARTIER-IA platform is mainly focused on research tasks, image edition and AI algorithms application, the previous heuristics are not the best fit for this usability study.

For these reasons, the Nielsen's heuristics were the selected instrument to perform the heuristic evaluation on the CARTIER-IA platform. This set is composed of ten heuristics, which are listed below [18].

HR1: Visibility of system status.

HR2: Match between system and the real world.

HR3: User control and freedom.

HR4: Consistency and standards.

HR5: Error prevention.

HR6: Recognition rather than recall.

HR7: Flexibility and efficiency of use.

HR8: Aesthetic and minimalist design.

HR9: Help users recognize, diagnose, and recover from errors.

HR10: Help and documentation.

A total of six experts were involved in the heuristic evaluation. Four of these experts were HCI experts (web developers and researchers), and two of them both HCI experts and domain experts (a Ph.D.

student and clinical data scientist) [21]. In fact, these double experts had used the CARTIER-IA platform as users before performing the heuristic evaluation.

The heuristic evaluation was carried out using a template with guidelines to support the evaluation and issues' reporting. Each evaluator had only access to his/her own report, in order to avoid biases. The evaluation template had three fields to collect the evaluator's name, the name of the tool evaluated, and the browser that they employed to access the platform.

Finally, the template provided a table with three columns (heuristic name, score from 1 to 10 and problems detected) and one row per problem detected within each heuristic.

## III. The CARTIER-IA Platform

### A. Data Collection

As introduced in section I, one of the motivations of developing the platform is to unify data from different sources and arrange them into a more friendly structure. Due to this requirement, the CARTIER-IA platform provides two types of data upload processes.

First, a structured data uploader. The platform allows users to upload spreadsheets of data at different levels, containing information associated to patients, image studies or files. The platform also supports longitudinal structures (repeated measurements or data for the same patient over time).

In addition, data schemas are flexible to vary among different projects, so a project might contain a structured data schema completely different from another.

This flexibility is accomplished through the Django ORM (a database-abstraction API). The Django ORM API provides access to a relational database with the structure shown in Fig. 1. The platform is mainly organized through projects, which will hold data from different patients. Structured data is stored as JSON object at different levels (patient, study or file), which provides the support to modify the data schemas across projects.

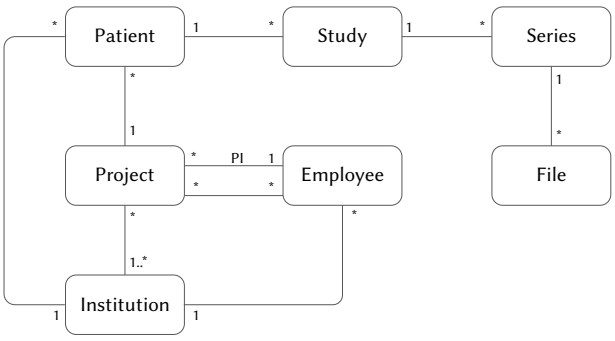



Fig. 1. Schematic overview of the platform's data structuration.

On the other hand, the platform provides a second data uploader; an image data uploader. This uploader allows users to upload a set of image studies through compressed files. The service handles the DICOM format, read some of the metadata tags, check the integrity of data anonymization of each file and validates the metadata against already stored structured data, linking them if applicable.

In this regard, imaging data relies on three entities: studies, series, and files. These entities provide the structure to manage data about the DICOM images that will be uploaded to the platform.

The image files are referenced through these entities and stored through file systems. The type of file system can be modified, allowing flexibility in the selection of the storage technology.

Finally, the platform integrates an external tool to make the data management more powerful. A REDCap instance is connected to the platform to enable the importation of its data, thus providing another layer of data unification. REDCap (Research Electronic Data Capture) is an electronic data capture (EDC) software and workflow methodology for creating and designing clinical research databases [22].

### B. Image Edition

Another main functionality of the platform is its image editor. When DICOM images are uploaded, users can explore them more closely through this tool. One of the benefits of this image editor is that is fully integrated within the platform, so it is not necessary to use external tools to carry out image modifications.

This is possible because, as explained in section II, image edition takes place in the browser through the Cornerstone.js (https://github.com/cornerstonejs/cornerstone) framework, an open-source library to parse and render DICOM files.

Thanks to this approach, users can edit the images they are currently exploring, and decide later if they can make these annotations and modifications persistent.

These modifications are not stored along the image itself, but as JSON objects containing all the necessary meta-data regarding the carried-out modifications or annotations. By storing the modifications as standalone objects, it is possible to explore the annotations made by other users, compare them against each other or even to have a version control of the modifications on each image.

The majority of image edition tools and functionalities are supported by another open-source framework, which provides an extensible solution for creating tools on top of Cornerstone.js (https://github.com/cornerstonejs/cornerstoneTool). Specifically, the following tools are available through the image editor:

- Brush and scissors tools for image segmentation
- Segmentation layers and brush size selectors to ease the segmentation process of the images
- Length and area tools to measure image fragments
- Annotation tools
- Zoom tools
- A crop tool (computed on the backend)
- A tool to apply uploaded and validated AI scripts (computed on the backend)

### C. Artificial Intelligence Support

The feature in which this paper is focused is the Artificial Intelligence integration within the platform. This feature has two main motivations:

- To offer the benefits from Artificial Intelligence algorithms in situ, without the necessity of leaving the platform to applying these algorithms
- To provide a friendly interface to apply AI scripts and open their use to non-specialized users

This feature allows researchers to upload their AI scripts into the platform and make them available to other users. Only researchers with privileges can add new scripts, which need to be thoroughly tested by the corresponding researcher before integrating them into the platform to ensure a reliable functionality.

The platform also provides an uploader to define the algorithm's meta-data. In this case, algorithms' meta-data is highly important to properly integrate the scripts within the platform. These meta-data provide information about the algorithm's output (a modified image, a set of measures, a segmentation mask, etc.), its applicability (as their

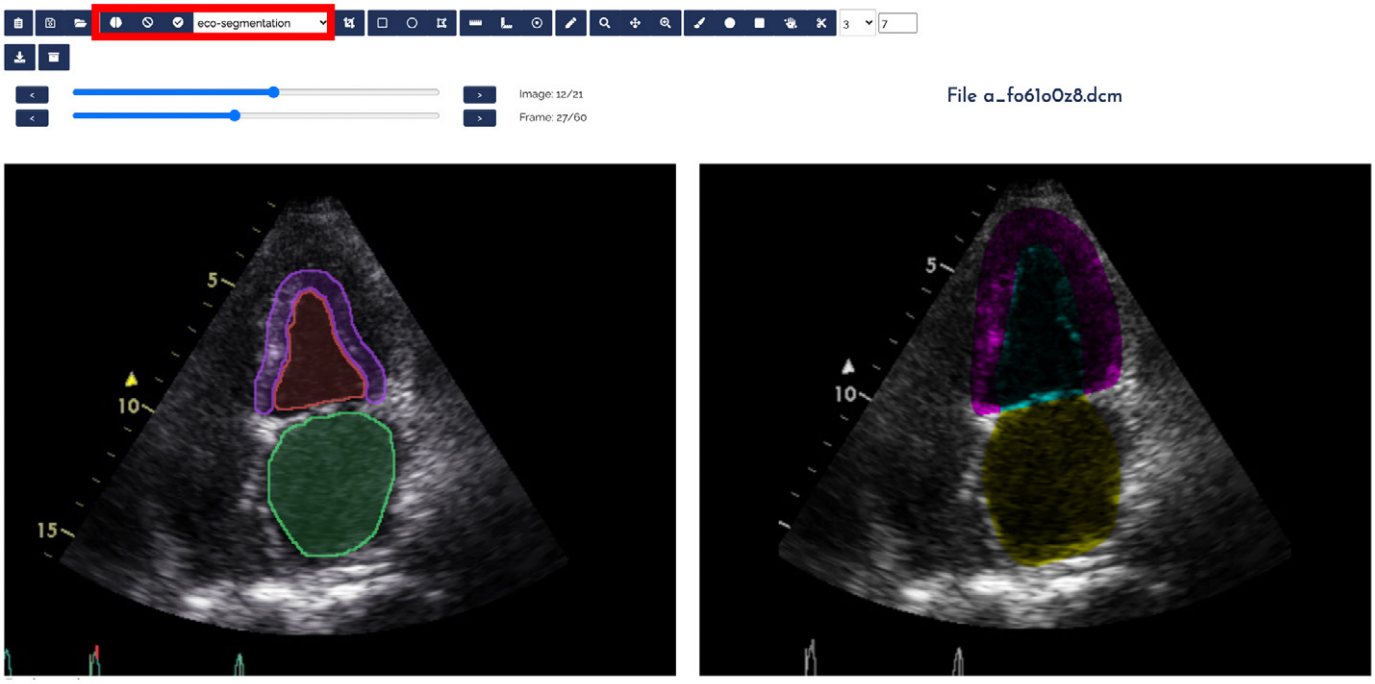



Fig. 2. Screenshot of a manual segmentation (left) and the AI algorithm output (right).

application might be limited to specific DICOM modalities) or other parameters depending on the output.

It is important to clarify that the algorithms need to be pre-trained before their integration into the platform. For this reason, the uploader also provides a field to upload the exported model or the models' weights depending on the type of AI algorithm employed.

To sum up, to integrate an algorithm into the platform it is necessary to provide the pre-trained model, and the script that makes use of the pre-trained model with the goal of enabling their invocation by the platform's AI module.

Once an algorithm has been integrated, it will be available at image editor. To apply an algorithm, the user just needs to click the AI button and select one of the available scripts for the current image being displayed. When the user confirms the application, the platform will yield the result which, depending on the algorithm's output type, could result in displaying a new image, an inferred diagnosis or the addition of AI-driven measurements as new structured data.

## IV. Use Cases

This section provides two application uses of the AI integration within the CARTIER-IA platform as an example of how the platform behaves when dealing with different types of AI algorithms.

### *A. Manual vs. Artificial Intelligence Segmentation*

Segmentation of medical images is a relevant procedure within the field of medical image processing. Its ultimate goal is to identify different elements and features in medical images to detect abnormalities or other characteristics of interest.

For this reason, one of the most relevant features of the image viewer is the possibility of performing the segmentation of the stored DICOM images in place.

As explained in the previous section, the image viewer relies on different tools to provide a complete set of image processing functionalities. Among them, the platform offers different brushes to perform image segmentations manually and store them as JSON objects that can be retrieved and further processed.

But along with the manual segmentation, researchers can integrate deep learning models whose outputs are automatically generated segmentations. To do that, users can select among the available AI algorithms in the platform and simply confirm their choice (red rectangle in Fig. 2). The algorithm choice is processed in the back end, which consults the algorithm's meta-data and, depending on the output, performs different actions. In this case, the output is a segmented image, so this result is sent back to the client and displayed in the viewer next to the original image (Fig. 2).

This interface organization allows users to compare their manual segmentation with the algorithm's result, which could provide new information or assist the user with their own image segmentation. And it can also be used in the reverse scenario. If it is a trained physician or technician who performs the manual segmentation, this interface could be used to improve the artificial intelligence algorithm by active learning.

### *B. Measurements*

The application of AI scripts is not only limited to image segmentation. As explained throughout this work, the platform manages both imaging data and structured data. Structured data also provides crucial information regarding patients and their monitoring, diagnoses, treatments, etc.

In this context, the CARTIER-IA platform also supports the execution of AI algorithms that, based on the input image or even input structured data, return a dataset containing new inferred information. The algorithm's results are persistently stored along with the rest of the patient's, study's or file's structured data, making them available for other users when exploring the project.

The process to apply these kinds of algorithms is exactly the same. However, in this case, instead of returning a new image, the back end executes the algorithm, stores the newly generated variables and sends a confirmation to the client (Fig. 3).

After the confirmation, users can see the measures yielded by the algorithm at the specified level. Fig. 4 shows a new variable generated by the algorithm "ai_DummyECO-script", which is stored under a new


Fig. 3. Screenshot of AI algorithms' measurements output.

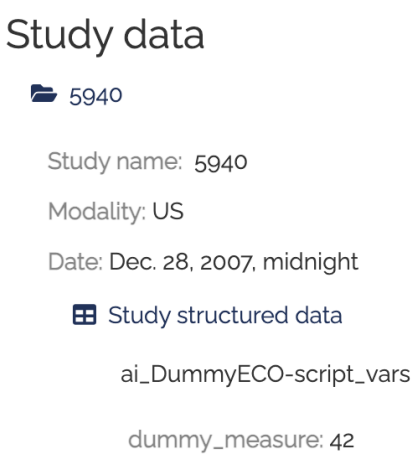



Fig. 4. Automatically generated variables after applying a measurement AI algorithm.

category with the format "<algorithm_name>_vars" to differentiate them from the original study variables.

## V. Heuristic Evaluation Results

Each expert was identified by a number (E1, E2, E3, E4, E5, E6) in order to present the outputs of the heuristic evaluation. The heuristic evaluation was performed on the whole platform, but only the DICOM editor and AI tool-related issues are being discussed given the focus of this work.

Fig. 5 shows the total average value assigned to the problems identified under each heuristic. Values close to 1 indicate that experts detected non-relevant issues, and values close to 10 implies that the issues are relevant and severe. A zero value represents that experts did not identify any problems in that heuristic. Not only is the severity of the problems important, but also the absolute number of issues to solve in each heuristic (Fig. 6).

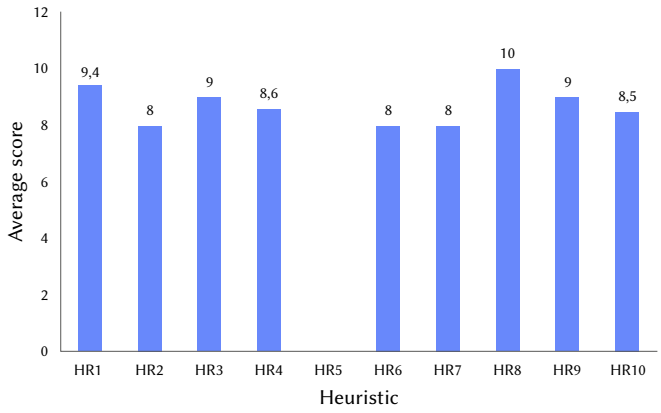



Fig. 5. Average score for each heuristic rule regarding the DICOM editor and AI tool.

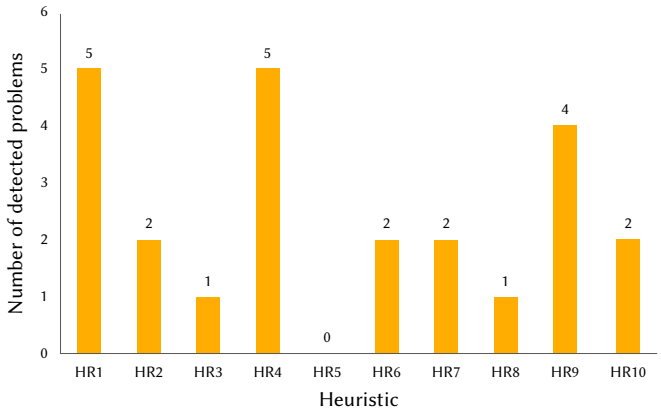



Fig. 6. The total number of detected problems regarding the DICOM editor and AI tool per Nielsen's heuristic.

The heuristics with the largest number of usability issues identified for the DICOM viewer and AI tool were HR1 (*Visibility of system status*) and HR4 (*Consistency and standards*), both with 5 problems

detected. The average severity scores for these heuristics are 9.4 and 8.6, respectively.

The HR1 problems are mainly related to the absence of progress bars and the actions that can take place within the image editor, which are not clearly explained.

Regarding HR4-related issues, they are focused on the correctness of the metaphors used for the icons that represent each functionality button. In addition, some misfunctions on these image processing functionalities were also identified.

On the other hand, there are three heuristics that also obtained high severity ratings: HR3 and HR9 with a score of 9 and HR8 with a score of 10.

Regarding HR8 (*Aesthetic and minimalist design*), this high score is due to the fact that only one issue was encountered within this category (also related with the DICOM editor's icons), but E6 assigned a score of 10 because of its relevance. Specifically, this issue pointed out the great quantity of icons employed for the editor toolbar and their difficulty to clearly convey their meaning.

Only one issue was identified under HR3 (*User control and freedom*) but also with a high severity rating (9). In this case, E6 identified the impossibility of undo or redo actions taken place within the tool.

Four issues were identified in the HR9 (*Help users recognize, diagnose, and recover from errors*), obtaining a score of 9, too. In this case, experts identified the lack of information when an AI script or a DICOM image fails and the impossibility of recovering from this kind of errors. This heuristic also includes the issue that the DICOM editor does not support the reset of the modifications made on the images.

Finally, a lower number of issues were encountered in the rest of heuristics:

- Better explanations regarding the editor's functionalities (HR2 and HR6)
- Better explanations regarding the results yielded by the functionalities supported on the editor (HR6)
- Keyboard shortcuts and AI integration for advanced users (HR7)
- Lack of documentation, specifically regarding the AI tool, which could be complex to understand (HR10)

## VI. Discussion

The heuristic evaluation identified different design issues regarding the image editor and AI algorithms' application. These usability evaluations are crucial to iteratively provide more robust and friendly interfaces to perform complex tasks such as the ones supported by the CARTIER-IA platform.

The results derived from the heuristic evaluation shown very high scores. This is due the great relevance that experts gave to usability in the image editor tool. The image editor is a powerful component of the CARTIER-IA platform, because not only provides edition functionalities, but also is the integration point for applying AI algorithms. For these reasons, offering good user experience in the image editor interface is crucial, and thus every usability issue encountered has high relevance.

The majority of issues were related to the toolbar, which relies on several icons to depict the supported functionalities. However, these icons were not very clear to the experts. In addition to this topic, some experts also pointed out the necessity of explaining the functionalities more thoroughly, especially the AI algorithms.

In this version of the platform, algorithms are listed in the interface and the user can apply them directly. However, only the names of the algorithms are displayed, which can be confusing, as the algorithm's name could provide little or no information at all regarding its outputs and results.

One of the design parameters for the AI integration was to make the application process straightforward both for skilled and non-skilled users. However, simplifying too much this process can also have drawbacks. Non-skilled users could question the algorithms' outputs if there are no further explanations regarding the process nor the interpretation of the results, because the AI tool works as a black box in its current version.

It is crucial to find balance between implementing a simple interface but also displaying enough information to understand the actions carried out within the platform.

Providing user-friendly interfaces in the health domain could make convoluted tasks more straightforward and thus, save time for physicians. As it has been shown, this kind of interfaces could also bring closer tasks for which users are not specialized nor trained (such as AI algorithms programming).

Another important benefit from integrating AI algorithms in a medical data management platform is that the trained models can be improved. Although in its current version the platform only provides an interface for executing AI algorithms (because models need to be pre-trained before their integration into the platform), this approach sets the foundations for future improvements, including the possibility of training AI models directly from the platform.

For example, manual segmentations can be carried-out by the users within the platform, which results in new data to train the existing models. On the other hand, users can also label the algorithms' outputs depending on their performance, thus laying the foundations for improving the models through active learning.

On the other hand, there is room for improvement regarding the algorithms' validation. Currently, researchers are responsible of the validation of their scripts, but another validation layer can be implemented to analyze and test these scripts automatically before carrying out the integration. The metrics obtaining from the testing of the scripts could complement the information of each algorithm to generate more confidence regarding the platform's AI support.

Finally, we want to mention that a heuristic evaluation does not ensure identifying all the problems that could affect a real user in a real context while using the platform. There are studies that point out that the problems detected by experts are not necessarily the actual problems that will affect the end users of the platform [23]. To alleviate this limitation, we included one expert with extensive knowledge of the platform's domain, although subsequent research will explore usability from the researchers and physicians' point of view.

## VII. Conclusions

This paper presents a collaborative platform for the management of medical data and imaging. The platform has several features to provide support for a variety of functionalities, such as a DICOM viewer and editor. Among these tools there is the possibility of integrating artificial intelligence scripts to make the application process straightforward to non-specialized users.

Given the implication of all these features within the platform, a heuristic evaluation has been carried out to identify usability issues of the DICOM viewer and AI algorithms' integration in the current version of the platform. This evaluation gives hints on the aspects that need to be improved to provide better user experience to researchers and physicians.

Future research lines will involve the resolution of every usability issue identified, as well further usability tests including other techniques such as the PSSUQ questionnaire or usability labs.

## Acknowledgment

This research work has been supported by the Spanish *Ministry of Education and Vocational Training* under an FPU fellowship (FPU17/03276). This work was also supported by national (PI14/00695, PIE14/00066, PI17/00145, DTS19/00098, PI19/00658, PI19/00656 Institute of Health Carlos III, Spanish Ministry of Economy and Competitiveness and co-funded by ERDF/ESF, "Investing in your future") and community (GRS 2033/A/19, GRS 2030/A/19, GRS 2031/A/19, GRS 2032/A/19, SACYL, Junta Castilla y León) competitive grants.

## References

[1] A. Rajkomar, J. Dean, and I. Kohane, "Machine Learning in Medicine," (in eng), *N Engl J Med,* vol. 380, no. 14, pp. 1347-1358, 04 2019, doi: 10.1056/NEJMra1814259.

[2] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," (in eng), *Med Image Anal,* vol. 42, pp. 60-88, Dec 2017, doi: 10.1016/j.media.2017.07.005.

[3] S. González Izard, R. Sánchez Torres, Ó. Alonso Plaza, J. A. Juanes Méndez, and F. J. García-Peñalvo, "Nextmed: Automatic Imaging Segmentation, 3D Reconstruction, and 3D Model Visualization Platform Using Augmented and Virtual Reality," (in eng), *Sensors (Basel),* vol. 20, no. 10, p. 2962, 2020, doi: 10.3390/s20102962.

[4] S. G. Izard, J. A. Juanes, F. J. García Peñalvo, J. M. G. Estella, M. J. S. Ledesma, and P. Ruisoto, "Virtual Reality as an Educational and Training Tool for Medicine," *Journal of Medical Systems,* vol. 42, no. 3, p. 50, 2018/02/01 2018, doi: 10.1007/s10916-018-0900-2.

[5] X. Liu *et al.*, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis," *The lancet digital health,* vol. 1, no. 6, pp. e271-e297, 2019.

[6] A. Vázquez-Ingelmo, F. J. García-Peñalvo, and R. Therón, "Information Dashboards and Tailoring Capabilities - A Systematic Literature Review," *IEEE Access,* vol. 7, pp. 109673-109688, 2019, doi: 10.1109/ACCESS.2019.2933472.

[7] A. Vázquez-Ingelmo, F. J. García-Peñalvo, R. Therón, D. A. Filvà, and D. F. Escudero, "Connecting domain-specific features to source code: towards the automatization of dashboard generation," *Cluster Computing. The Journal of Networks, Software Tools and Applications,* p. In Press, 2020, doi: 10.1007/s10586-019-03012-1.

[8] A. Sarikaya, M. Correll, L. Bartram, M. Tory, and D. Fisher, "What Do We Talk About When We Talk About Dashboards?," *IEEE Transactions on Visualization Computer Graphics,* vol. 25, no. 1, pp. 682 - 692, 2018.

[9] Y. Zhonggen, "Visualizing Artificial Intelligence Used in Education Over Two Decades," *Journal of Information Technology Research (JITR),* vol. 13, no. 4, pp. 32-46, 2020, doi: 10.4018/JITR.2020100103.

[10] J. C. Sánchez-Prieto, J. Cruz-Benito, R. Therón Sánchez, and F. J. García Peñalvo, "Assessed by Machines: Development of a TAM-Based Tool to Measure AI-based Assessment Acceptance Among Students," *International Journal of Interactive Multimedia and Artificial Intelligence,* vol. 6, no. 4, p. 80, 2020.

[11] O. Bernard *et al.*, "Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved?," *IEEE Trans Med Imaging,* vol. 37, no. 11, pp. 2514-2525, Nov 2018, doi: 10.1109/TMI.2018.2837502.

[12] P. Mildenberger, M. Eichelberg, and E. Martin, "Introduction to the DICOM standard," *European radiology,* vol. 12, no. 4, pp. 920-927, 2002.

[13] D. S. Marcus, T. R. Olsen, M. Ramaratnam, and R. L. Buckner, "The Extensible Neuroimaging Archive Toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data," (in eng), *Neuroinformatics,* vol. 5, no. 1, pp. 11-34, 2007, doi: 10.1385/ni:5:1:11.

[14] E. Gibson *et al.*, "NiftyNet: a deep-learning platform for medical imaging," (in eng), *Comput Methods Programs Biomed,* vol. 158, pp. 113-122, May 2018, doi: 10.1016/j.cmpb.2018.01.025.

[15] A. García-Holgado and F. J. García-Peñalvo, "Preliminary validation of the metamodel for developing learning ecosystems," in *Fifth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'17) (Cádiz, Spain, October 18-20, 2017)* J. M. Dodero, M. S. Ibarra Sáiz, and I. Ruiz Rube Eds., (ACM International Conference Proceeding Series (ICPS). New York, NY, USA: ACM, 2017.

[16] A. García-Holgado and F. J. García-Peñalvo, "Validation of the learning ecosystem metamodel using transformation rules," *Future Generation Computer Systems,* vol. 91, pp. 300-310, 2019, doi: 10.1016/j.future.2018.09.011.

[17] Django Software Foundation. "Django Web Framework." https://www.djangoproject.com/ (accessed 15/03/2015.

[18] J. Nielsen, "Heuristic evaluation," in *Usability inspection methods*, vol. 17, J. Nielsen and R. L. Mack Eds., no. 1): John Wiley & Sons, Inc., 1994, pp. 25-62.

[19] A. Tarrell, L. Grabenbauer, J. McClay, J. Windle, and A. L. Fruhling, "Toward improved heuristic evaluation of EHRs," *Health Systems,* vol. 4, no. 2, pp. 138-150, 2015/07/01 2015, doi: 10.1057/hs.2014.19.

[20] D. Armijo, C. McDonnell, and K. Werner, *Electronic health record usability: Evaluation and use case framework. AHRQ Publication No. 09(10)-0091-1-EF.* Rockville, MD: Agency for Healthcare Research and Quality, 2009.

[21] J. Nielsen, "Finding usability problems through heuristic evaluation," in *CHI '92: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* New York, NY, USA: ACM, 1992, pp. 373–380.

[22] P. A. Harris, R. Taylor, R. Thielke, J. Payne, N. Gonzalez, and J. G. Conde, "Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support," *Journal of biomedical informatics,* vol. 42, no. 2, pp. 377-381, 2009.

[23] R. Khajouei, A. Ameri, and Y. Jahani, "Evaluating the agreement of users with usability problems identified by heuristic evaluation," *International Journal of Medical Informatics,* vol. 117, pp. 13-18, 2018/09/01/ 2018, doi: https://doi.org/10.1016/j.ijmedinf.2018.05.012.

Francisco José García-Peñalvo

He received the degrees in computing from the University of Salamanca and the University of Valladolid, and a Ph.D. from the University of Salamanca (USAL). He is Full Professor of the Computer Science Department at the University of Salamanca. In addition, he is a Distinguished Professor of the School of Humanities and Education of the Tecnológico de Monterrey, Mexico. Since 2006 he is the head of the GRIAL Research Group GRIAL. He is head of the Consolidated Research Unit of the Junta de Castilla y León (UIC 81). He was Vice-dean of Innovation and New Technologies of the Faculty of Sciences of the USAL between 2004 and 2007 and Vice-Chancellor of Technological Innovation of this University between 2007 and 2009. He is currently the Coordinator of the PhD Programme in Education in the Knowledge Society at USAL. He is a member of IEEE (Education Society and Computer Society) and ACM.

Andrea Vázquez-Ingelmo

Andrea Vázquez-Ingelmo received the bachelor's degree in computer engineering from the University of Salamanca, Salamanca, in 2016 and the master's degree in computer engineering from the same university in 2018. She is a member of the Research Group of Interaction and eLearning (GRIAL), where she is pursuing her PhD degree in computer sciences. Her area of research is related to human-computer interaction, software engineering, information visualization and machine learning applications.

Alicia García-Holgado

She received the degree in Computer Sciences (2011), a M.Sc. in Intelligent Systems (2013) and a Ph.D. (2018) from the University of Salamanca, Spain. She is member of the GRIAL Research Group of the University of Salamanca since 2009. Her main lines of research are related to the development of technological ecosystems for knowledge and learning processes management in heterogeneous contexts, and the gender gap in the technological field. She has participated in many national and international R&D projects. She is a member of IEEE (Women in Engineering, Education Society and Computer Society), ACM (and ACM-W) and AMIT (Spanish Association for Women in Science and Technology).

Jesús Sampedro-Gómez

Jesús Sampedro-Gómez is industrial engineer by the Universidad Politécnica of Madrid. He is also a PhD student by the University of Salamanca and works as data scientist in the Cardiology Department of the University Hospital of Salamanca.

Antonio Sánchez-Puente

Antonio Sánchez Puente, PhD, is a junior researcher from CIBER working at the cardiology department of the University Hospital of Salamanca as a data scientist. He was awarded his doctorate in physics by the University of Valencia for his study of gravity theories before turning his career around the application of artificial intelligence in medicine.

Víctor Vicente-Palacios

Víctor Vicente-Palacios holds a PhD from the University of Salamanca in Statistics and works as a Data Scientist at Philips Healthcare in the area of AI applied to Medicine. He is also an alumnus of the Data Science for Social Good program (University of Chicago) and organizer of PyData Salamanca.

P. Ignacio Dorado-Díaz

P. Ignacio Dorado-Díaz holds a PhD from the University of Salamanca in Statistics. He is currently working as a research coordinator in the Cardiology Department of the Hospital de Salamanca in addition to being a professor at the Universidad Pontificia de Salamanca.

Pedro L. Sánchez

Pedro Luis Sánchez holds a doctorate in medicine from the University of Salamanca. He is currently the head of the Cardiology Department at the University Hospital of Salamanca, in addition to being a professor at the University of Salamanca.

# The Application of Artificial Intelligence in Project Management Research: A Review

Jesús Gil Ruiz[1]*, Javier Martínez Torres[2], Rubén González Crespo[3]

[1] School of Doctorate Programs, Universidad Internacional de La Rioja, Logroño, La Rioja (Spain)
[2] Department of Applied Mathematics I, Universidad de Vigo, Vigo (Spain)
[3] School of Engineering and Technology, Universidad Internacional de La Rioja, Logroño, La Rioja (Spain)

Received 11 May 2020 | Accepted 7 October 2020 | Published 18 December 2020

## Abstract

The field of artificial intelligence is currently experiencing relentless growth, with innumerable models emerging in the research and development phases across various fields, including science, finance, and engineering. In this work, the authors review a large number of learning techniques aimed at project management. The analysis is largely focused on hybrid systems, which present computational models of blended learning techniques. At present, these models are at a very early stage and major efforts in terms of development is required within the scientific community. In addition, we provide a classification of all the areas within project management and the learning techniques that are used in each, presenting a brief study of the different artificial intelligence techniques used today and the areas of project management in which agents are being applied. This work should serve as a starting point for researchers who wish to work in the exciting world of artificial intelligence in relation to project leadership and management.

## Keywords

Artificial Intelligence, Decision Support Systems, Evolutionary Diffuse Hybrid Neuronal Network, Project Management, Project Success, Critical.

DOI:10.9781/ijimai.2020.12.003

## I. Introduction

In recent decades, projects have tended to increase in complexity to the point where they have become mega projects such as, for example, the particle accelerator (CERN) or the photovoltaic plants (BEN BAN solar) with the power of almost two nuclear reactors (1.8 GW). Meanwhile, the attendant industrial growth has resulted in a greater degree of competence when addressing these projects in terms of their control and development, which has become a necessity since the projects often involve extremely tight profit margins. Adopting certain project management methodologies (e.g., PMI, [130], IPMA, and PRINCE) allows us to manage the start and the evolution of a project in the most optimal way possible, controlling and responding to any problems that arise during the project, facilitating their completion and approval before any further risks arise. However, these methodologies are arguably not sufficient since the processes must be clearly structured with complete and clear control of the project in all the relevant areas. The aim must be to improve the experience of the project manager when dealing with the various adverse situations that will likely be encountered in the development of the project while simultaneously preventing errors due to a lack of planning or management, such as in portfolio management [41]. While the desired project management methodology (PMP) practices are currently being implemented – which allow for the best possible management of a project – as noted above, the processes must be clearly structured [142] and all areas of the project must be tightly controlled, including in terms of the information systems [66].

In fact, the current methodologies are largely insufficient since the project manager is generally left to deal with the decision making, who, based on his or her professional experience, must make "intuitive" decisions based on previous cases when facing a problem with infinite variables and possibilities. Here, it is virtually impossible to face all the issues and challenges that today's projects entail. In fact, there are a number of diverse reasons why projects tend to fail. However, after more than ten years working on projects and learning about other professionals' experiences, we would highlight the following:

- Unassembled objectives or objectives that are not clearly defined.
- There is no communication protocol.
- Lack of definition of roles and responsibilities.
- Expectation management.
- Scope Corruption.
- Ignore Project Risks.
- Lack of involvement of participants.
- Absence of formal planning.
- Estimated errors / unrealistic.
- Absence of methodologies, templates and documentation.
- Lack of resources.
- Absence of evidence or little focus on quality.
- Little formalized modification process.
- Lack of training.
- Little or no address support.

* Corresponding author.
E-mail address: jesus.gil@unir.net

While all these points can be improved with a clear PMP, they will always depend on the human factor, and many of them are difficult to deal with, even for an experienced project manager. In view of this, artificial intelligence (AI) can play an important role in a variety of areas.

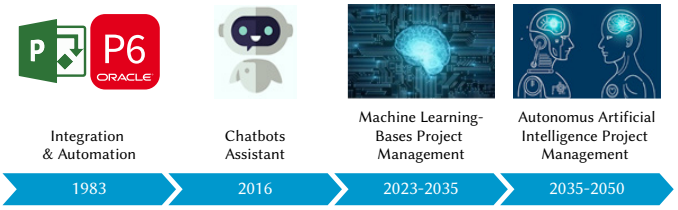



Fig. 1.Evolution of AI in project management [100].

Fig. 1 shows the evolution that has taken place in the last 37 years, and what is expected in the future.

### *Integration, Automation, and Chatbot Assistants*

The first phase involved the integration of task automation software such as Microsoft Project and Primavera (Oracle), which first appeared in 1983. In recent years, chatbot assistants are being used for meetings and management equipment recaps and reminders, etc. While in everyday life, we have been surrounded by chatbots for several years, the area is still in its infancy in the world of project management.

### *Project Management Based on Machine Learning*

The third stage began with the purest concept of AI. In the area of project management, machine learning [132] has been implemented to allow for predictive and corrective analysis aimed at providing the project manager with data for decision making in terms of, for example, how to plan and manage project resources within certain parameters and restrictions or how to deal with problems and risks in order to achieve project success based on the history of past projects. In less than ten years, AI could work with the lessons learned from the project history and could suggest new project schedules, adapting [87] to the real time according to the performance of the resources and the progress of the project. An AI system could even alert the project manager about any possible risks and opportunities through the use of real-time project data analysis. A new vision will be created when it comes to directing projects by minimizing the risks involved in decision making. An AI system may be capable of making decisions for itself, which will herald the new era of AI [19], one that will mark the fourth phase of the evolution of project management.

The objective of this work is to review the new proposals emerging in the field of AI in the various areas, and to ascertain which techniques could be the most effective for ensuring the success of the projects. We also look at all the applications and uses [112] of AI in the broad field of project management, from the commercial development phase to the construction and commissioning phase, including its application in the areas of operation and maintenance. Numerous international studies have recently emerged in relation to optimization techniques such as neural networks [27], support vector machines [8], evolutionary algorithms [61], and hybrid systems [32] [2]. Given their relevance to PMP, these techniques will improve the experience of the project manager when facing the various adverse situations that will be encountered in the development of the project, and will help to prevent the errors resulting from a lack of planning or management.

## II. Summary of Machine Learning Techniques

A project has traditionally been classified as successful if it has complied with the following restrictions: scope, budget, and schedule.

The objective of this document is to review the new proposals related to AI to improve the success of the project and to ascertain the applications and uses that AI has in the broad field of project management, from the development phase of the business [91] to its start-up [154] and onto its operation [125] [165] and maintenance [95] [106] [156].

### *A. Individual Techniques*

We begin by outlining each of the techniques used in the field of project management.

### *1. Artificial Neural Networks (ANN)*

Neural networks attempt to simulate [162] the way the human brain works as closely as possible, and are currently used in a number of fields, including medicine, engineering, and construction management [120]. The neural network conforms to data patterns and offers better results. This is achieved through learning the network [165] and comparing the results of the neural network with the data of other projects until the performance of the neural network is optimized.

Neural networks have the following advantages [109]:

- The storage of information throughout the network.
- The ability to work with incomplete knowledge.
- Fault tolerance.
- The ability to carry out machine learning.
- A parallel-processing capacity.

These advantages make its implementation in computational models highly interesting in all fields of research, text analytics [119], and project management [76].

### *2. Neural Networks of High Order (HONNS)*

HONNs were originally proposed in the 1960s to perform nonlinear discrimination but were discarded due to the enormous amount of higher-order terms [43]. Beginning in the mid-1990s, several researchers relied on HONNs rather than ONNs to resolve specific classification problems [79]. In a high-order neuronal, the neuron outputs are fed back to the same neuron or to neurons in the previous layers, as shown in Fig. 2. The signals are transmitted in forward and backward directions. High-order artificial neural networks are mainly based on the Hopfield model.

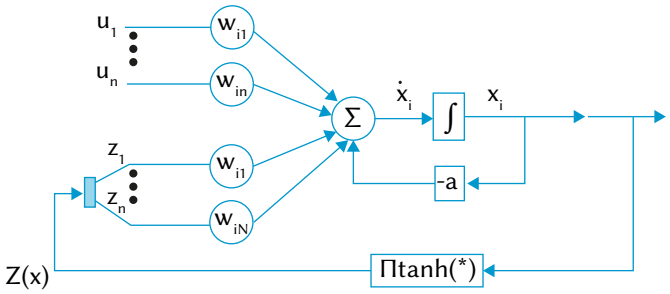



Fig. 2. Neural Network of high order [135].

### *3. Hopfield Neural Network (HNN)*

The HNN [146] is a form of high-order artificial neural network with a single layer of fully connected neurons (i.e., all neurons are also connected to each other, as shown in Fig. 3) and provides a method to resolve combinatorial optimization problems. A HNN is guaranteed to converge to a local minimum if a problem can be described as an energy function with a minimum corresponding to the optimal solution [60] [122].

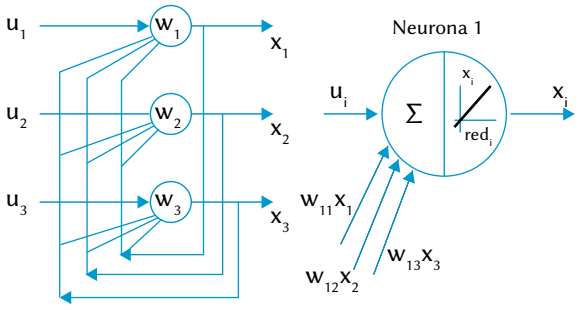



Fig. 3. Topology of Hopfield networks, here with 3 neurons as an example [102].

## 4. Fuzzy Logic (FL)

FL was initially proposed as a tool to describe uncertainty and inaccuracy [163]. Since it mimics the higher-order mode in which the human brain makes decisions in the face of uncertainty or vagueness, FL provides an effective way for automated systems to describe highly complex [48], poorly defined, or difficult to analyze subjects. In general, FL is composed of a fuzzifier, a rule base, an inference engine, and a defuzzifier [145] as shown in Fig. 4. The FL approach involves a number of issues that have yet to be overcome [57], such as the configuration of the membership function, the determination of the composition operator, and the acquisition of fuzzy rules that are specific [152] to the application. While FL parameters can be determined using the experience and knowledge of experts, determining these parameters in the absence of such experts remains difficult, especially in terms of complex issues.

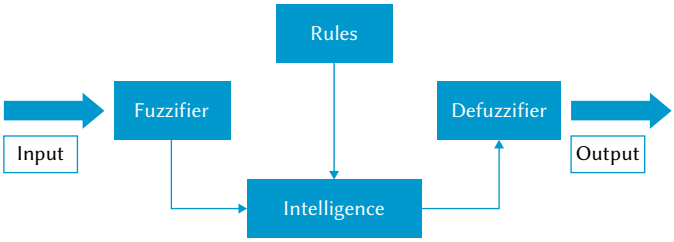



Fig. 4. Architecture of fuzzy logic systems [153].

## 5. Fuzzy Cognitive Maps (DCMs)

DCMs present an extension of cognitive maps and constitute a fuzzy graphical structure (as shown in Fig. 5) used to represent causal reasoning [96]. Their application is recommended for domains where the concepts and relationships are fundamentally fuzzy, such as politics, history, and strategic planning (projects) [51]. In the diagram shown in Fig. 5, each node represents a fuzzy set or an event that occurs to some degree. Here, it should be clarified that nodes are causal concepts and can model events, actions, values, objectives, or processes. Using this technique also provides the benefits of visual modelling, simulation, and prediction. Scenario analysis contributes to the identification of different alternatives to reach a future state [124]. This presents a flexible strategic planning method that is frequently used in technology management. While DCMs have been used for scenario analysis, there is a lack of methodologies and tools that allow for a fully effective quantitative analysis of the generated scenarios. In the area of information technology management [104], the simulation of software development projects and risk analysis in ERP maintenance stand out. While the use of DCMs has been proposed for the integration of strategic planning in relation to information systems and processes [136], the possible project options are neither represented nor analyzed. Furthermore, despite the DCM applications for the selection of information technology projects, the technique has not been linked to the organizational models that are obtained by describing the business architecture through business modelling activities.

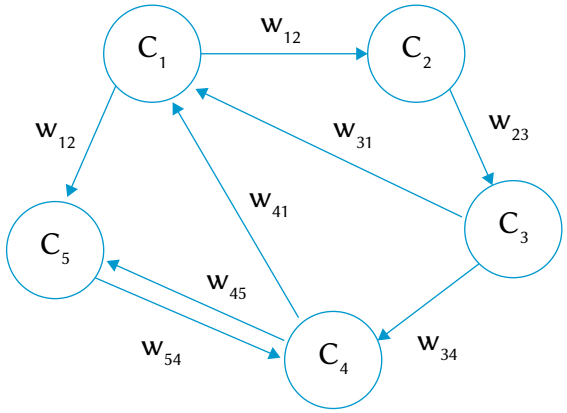


Fig. 5. Diffuse cognitive map topology [143].

## 6. Genetic Algorithms (GAs)

GAs present adaptive methods that can be used to resolve search and optimization problems and are based on the genetic process of living organisms. Over the generations, populations evolve in nature according to the principles of natural selection and the survival of the fittest, as postulated by Darwin (1859). The power of GAs lies in the fact that they present a robust technique and can successfully handle a wide variety of problems in different areas, including those where other methods encounter difficulties. While a GA is not guaranteed to find the optimal solution for a specific problem, empirical evidence suggests that solutions of an acceptable level can be identified in a timely manner when compared with other combinatorial optimization algorithms. The wide application of GAs is related to the problems for which there are no specialized techniques. In fact, these algorithms are used in countless applications, including in the fields of engineering [13], planning, games, and image processing [97]. Fig. 6 shows the working architecture of GAs.

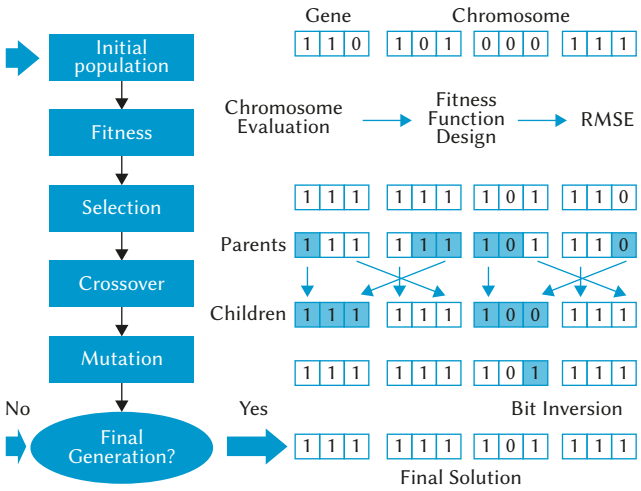



Fig. 6. Genetic Algorithms Diagram [70].

In general, the application of GAs to the planning of multiple projects that are to be executed simultaneously has yielded good results. In certain studies, a method based on penalties has been adopted [58] since it is difficult to obtain wholly correct solutions due to the complexity of the problem of optimization. While the identified solutions have, on the whole, been good, it is important to highlight

that, in some cases, the solutions lay outside of the algorithms, since the best solutions do not always meet all the restrictions of the problem.

### 7. Fast-Messy Genetic Algorithm (FmGA)

The fmGA [64] can efficiently identify optimal solutions to problems with a large number of permutations. This type of algorithm is known for its flexibility due to its capacity for being combined with other methodologies to obtain better results [160]. The difference between this and other genetic algorithms is based on the possibility of modifying building blocks [86] to identify the best partial solutions, which help us to focus on a faster global solution [65]. Fig. 7 shows the working architecture of Messy GA.

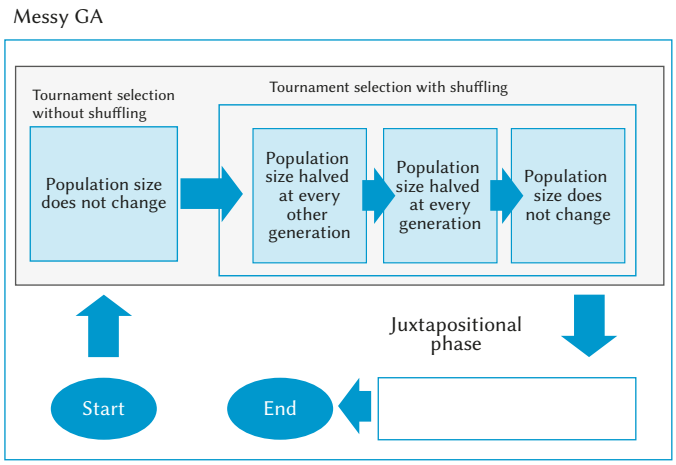



Fig. 7. Messy GA Architecture [99].

The algorithm is used in many applications, especially in relation to the resource management area of project management and civil engineering [45].

### 8. Support Vector Machine (SVM)

SVM presents a new form of learning, one that is more powerful than that using traditional learning tools. The technique can also be used to resolve data regression and categorization problems. Much like neural networks, SVM requires training and testing using a training dataset. The SVM functions allow for the better handling of unknown data and the technique generally has certain advantages over neural networks, often successfully applied to cost [10] and project management [158]. Within the area of classification, SVM belongs to the category of linear classifiers since it induces linear or hyperplane separators (as shown in Fig. 8), either in the original space of the input examples [20] – either separable or quasi-separable (noise) – or in a transformed space (characteristic space).

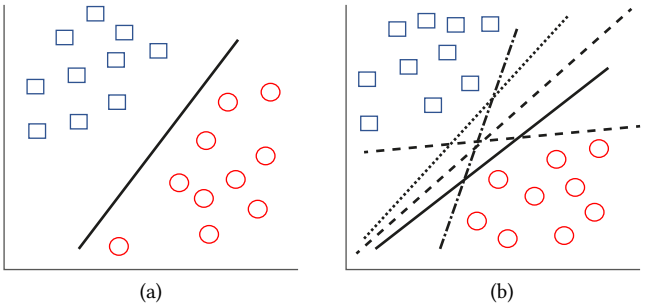



Fig. 8. Separation hyperplanes in a two-dimensional space of a set of separable examples from two classes: (a) example of separation hyperplane, (b) other examples of separation hyperplanes among the possible infinities [18].

### 9. Bootstrap Technique (BT)

The bootstrap method is a statistical technique used to estimate quantities across a specific population by averaging estimates from multiple small data samples [50]. Importantly, the samples are constructed by drawing observations from a large data sample one at a time before returning them to the data sample after they have been chosen. This allows a given observation to be included in a small sample more than once. This sampling approach is known as "replacement sampling." The bootstrap method can be used to estimate the size of a given population. This is achieved by repeatedly taking small samples, calculating the statistics, and then extracting the average. The bootstrap technique is a widely applicable and extremely powerful [159] statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method (e.g., ascertaining the probability that a project will be successful). This is achieved by training the model with a sample and evaluating the capacity of the model in relation to the samples not included in the main sample. A useful feature of the bootstrap method [17] is that the sample resulting from the estimates often forms a Gaussian distribution. This technique is used in a wide variety of sectors, including the fields of medicine, financial management [154] [111][142], and project management [68]. An example for risk analysis is shown in Fig. 9.

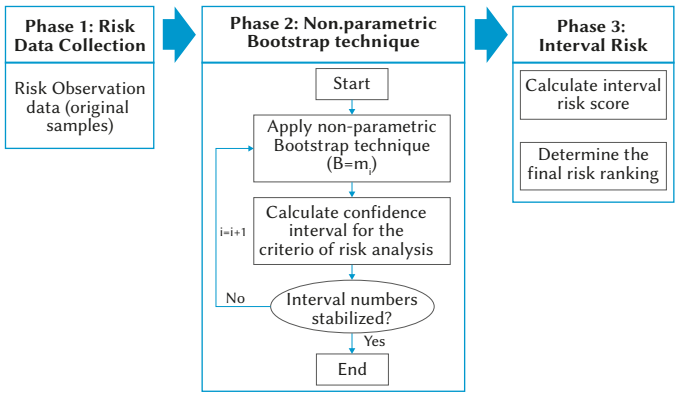



Fig. 9. Proposed approach for risk analysis [67].

### 10. K-Grouping Means

K-means presents an easy approach to creating groups of data from random datasets [84]. K-means grouping that incorporates heuristics such as Lloyd's algorithm is easy to implement, even in terms of large datasets, and has thus been widely used in many areas, such as market segmentation, computer vision, geostatistics, astronomy, and data mining in agriculture. The method is also used for the pre-processing in other algorithms, including in terms of identifying an initial configuration. While the main problem is that it cannot guarantee optimal convergence, it remains widely used due to its simplicity. Many algorithms can identify specific domains. K-means generally converges in practical applications [55], especially in pattern recognition problems. K-means clustering is also widely and commonly used due to its simplicity, while it does have certain inherent drawbacks, including having a fixed configuration for the optimal solution and being fairly time consuming.

### 11. Other Relevant Optimization Techniques

In the broad area of AI techniques, a number of well-known techniques are used, including the artificial bee colony algorithm [5], particle swarm optimization (PSO), and differential evolution (DE) [81]. There also exist various simple [128] or multi-objective Bayesian optimization algorithms [101].

### B. Hybrid Techniques

Here, we describe each of the hybrid techniques used in project management. These hybrid systems are the future of AI and automated project management.

#### 1. Neuro-Fuzzy (FNN)

The various logic and neural networks have special computational properties [4] that make them suitable for certain cases. For example, while neural networks offer advantages such as learning, adaptation, fault tolerance, parallelism, and generalization, they are not good at explaining how they have reached their decisions. In contrast, fuzzy systems – which reason using inaccurate information through an inference mechanism under linguistic uncertainty – are good at explaining their decisions but cannot automatically acquire the rules they use to make them. Meanwhile, neuro-diffuse systems [53] combine the learning capacity of RNAs with the linguistic interpretation power of diffuse inference systems. They are used in a multitude of applications and fields [137] [85], including mechanical engineering [155], image processing [74], electrical and electronic systems [129], forecasting and prediction [49], and risk identification in project management.

#### 2. Neural-Network-Adding Bootstrap

A bootstrap that adds neural networks presents a combination of multiple artificial neural network classifiers [151]. This method uses more than one ANN-based classifier, meaning the final decision is made from each classifier through a voting system. The model output is obtained as a linear combination of the experts' output and the combined weights are calculated based on the input. Bierman proposed a new method to aggregate multiple models using boot replicas of training data, which is known as "packaging". It has been shown that the generalizability of the model can be significantly improved through this approach. The "bagging" idea is used to build robust neural network models, or BAGNET models.

Rather than select a single neural network model, a BAGNET model combines several neural network models to improve the precision and robustness, as shown in Fig. 10 . The overall output of a BAGNET model presents a weighted combination of the outputs of individual neural networks. This approach has demonstrated a comparatively good performance.

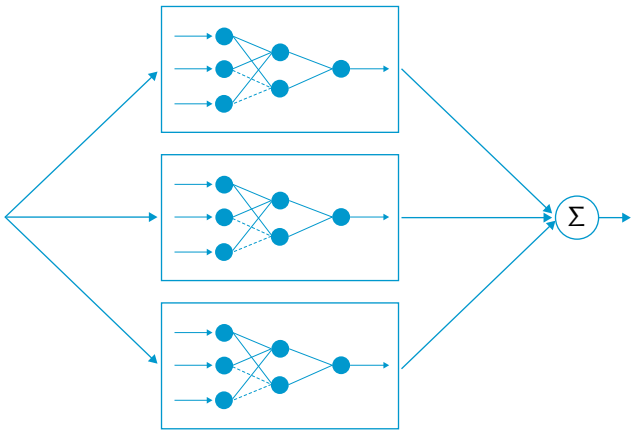



Fig. 10. Bagnet diagram [166].

#### 3. Neural Networks of Adaptive Reinforcement

The main difference between this method and the above method is that adaptive reinforcement neural networks [147] use weights that are readjusted in each iteration, affording less importance to the solutions that have not been correctly classified. As a result, the classifiers focus on more complex samples to obtain an increasingly faster solution. A number of interesting studies on this technique are currently available [126] [103].

#### 4. Fuzzy Rule-Based Systems (FRBS) and Genetic Fuzzy Systems (GFS)

FRBSspresent an extension of classical rule-based systems (hybrid systems, as shown in Fig. 11) [75] given that they deal with "IF-THEN" rules, the antecedents and consequents of which are made up of fuzzy logical statements, rather than classical ones. They have demonstrated their capacity for modelling, classification, and data mining problems in a large number of applications, which makes them highly useful for project management and control. A GFS is essentially a fuzzy system driven by a learning process based on evolutionary algorithms, which includes FL + GA, genetic programming, and evolution strategies, among other evolutionary algorithms [42].

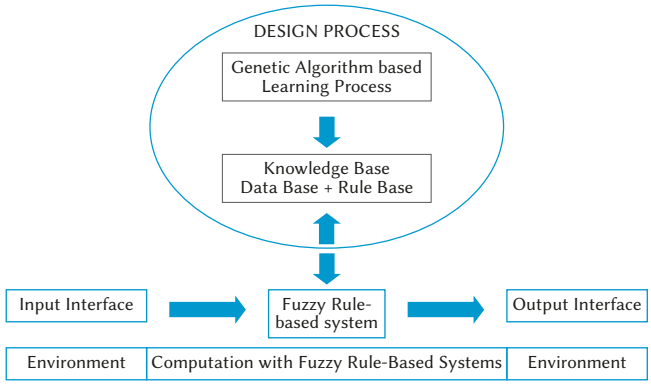



Fig. 11. FRBS Structure [71].

The central aspect of using a GA [40] for the machine learning of an FRBS is that the process can be analyzed as an optimization problem. This technique is frequently used in weather forecasts [46], the forecasting of renewable energy resources [144] (solar [90], wind [108]), military projects [52], and project management [12].

#### 5. Evolutionary Fuzzy Support Vector Machines Inference Model (EFSIM).

The inference model of evolutionary diffuse support vector machines (EFSIM) presents a hybrid technique [35] that incorporates three different AI techniques: FL, SVM, and fmGA, as shown in Fig. 12. In this hybrid system, the FL deals with any vagueness and approximate reasoning, the SVM acts as a supervisory learning tool to handle diffuse input–output mapping, and the fmGA functions to optimize the FL and SVM parameters. Interesting research on this technique has been conducted in relation to project management [33].

#### 6. Evolutionary Fuzzy Neural Inference Model (EFNIM)

EFNIM presents a resolution technique for hybrid systems [27] (composed of GA, FL, and NN, as shown in Fig. 13) that is used to resolve all types of problems. The complementary combination of its three elements maximizes the positive merits of each and helps to compensate for their inherent individual weaknesses. The GA is used for global optimization, the FL deals with uncertainties and handles approximate inferences, and the NN is used in the input–output mapping. Traditionally, the system has been used to resolve civil engineering problems [30] and presents a hybrid system that has great potential for assisting managers in implementing efficient long-term strategies and in taking the correct action for achieving the ultimate success of the project [94].

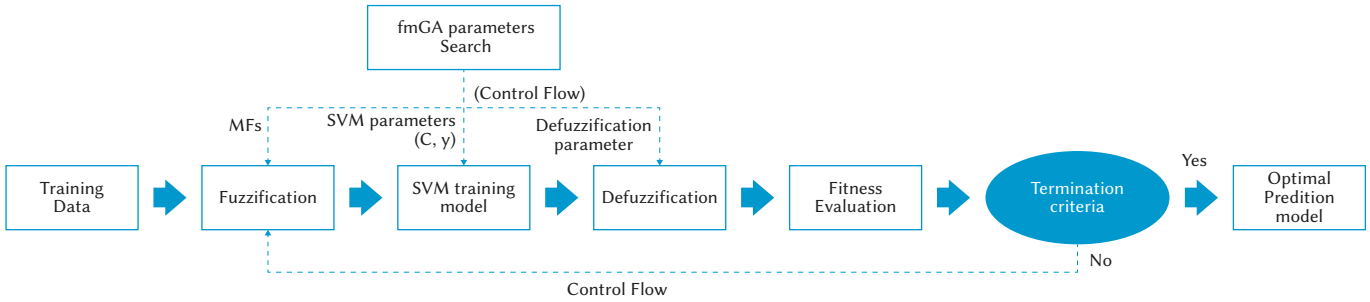



Fig. 12. Architecture of EFSIM [31].

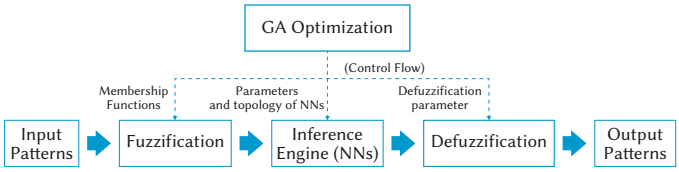



Fig. 13. EFNIM Architecture [26].

### 7. Evolutionary Diffuse Hybrid Neuronal Network (EFHNN)

The EFHNN mechanism is a fusion of HNN, FL, GA, and HNN. The advantage this system has over EFNIM is that the former is capable of handling deeper problems due to the large number of HNN models.

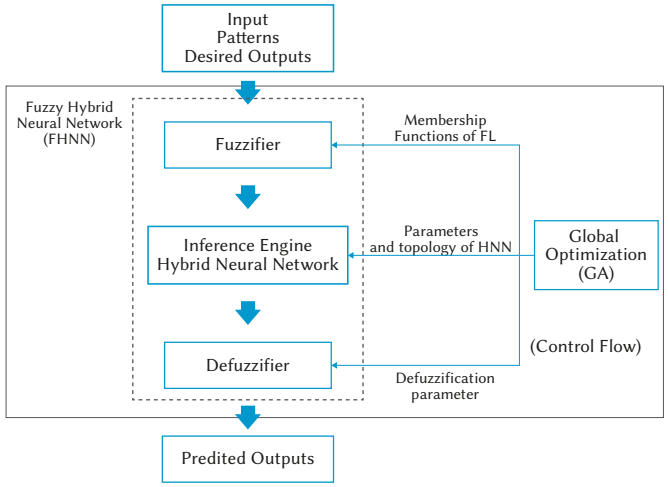



Fig. 14. EFHNN Architecture [29].

As noted, the proposed EFHNN for project management incorporates four AI approaches: NN, HONN, FL, and GA, as shown in Fig. 14. Here, the NN and HONN are composed of the inference engine, that is, the proposed HNN, the FL masters the fuzzifier and defuzzifier layers, and the GA optimizes the HNN and FL. Currently, there exist a small number of works that focus on this system in relation to project management in the field of civil engineering [32].

### 8. Other Relevant Optimization Techniques

Within the broad area of hybrid optimization techniques, there are a number that are worth mentioning. This includes firefly colony algorithm-based support vector regression (SAFCA-SVR) [37] and the fuzzy AHP and regression-based model [38]. Within the category of hybrid systems based on neural networks, there are a number of interesting examples, including multi-layer perceptron (MLP) combined with radial basis function network (RBFN) [77], diffuse object-oriented neural systems (OO-EFNIS) [92], wavelet-bootstrap-ANN (WBANN) [150], and neural networks combined with GA [113].

## III. Applications of Artificial Intelligence in Project Management

In the next section, we provide a brief description of the main studies that are being carried out in the field of AI in relation to project management. The table 1 is also provided, which presents the main authors working in each field along with the optimization techniques used in each study.

### A. Tenders

Tenders and technical offers that comprise the first phase of the project, wherein the initial estimates and designs are proposed in order to ascertain how much the project will cost as well as its scope. While the area is perhaps underexplored, there exist a number of interesting studies [38] on bidding strategies to support the decision making or AI models optimized to predict the project award price. In one study, the proposed model was used to analyze the data on bridge construction projects taken from the database of the Taiwan Public Construction Commission. The bid evaluation model and the cost probability curve model can be used as a strategic tool to quantify the project risks and to calculate the bids and tenders for construction projects. Another study [148] focused on machine learning and AI in terms of their impact on personal selling and sales management, with the impact discussed in relation to a small area of sales and research practice based on the seven steps of the sales process. From this, the implications for theory and practice can be derived.

### B. Project Health

A number of studies exist that focus on project management in relation to health. The studies are fairly diverse and include research [73] on achieving strategic control over the project's cash flows in order to develop appropriate strategies that apply factors such as the task execution time, the construction rate, and the demand for resources for cash-flow control. There are also a number of studies that analyze the project risks, with a model proposed [94] for risk analysis using the unrestricted automatic causality of data from various software projects. Here, it was demonstrated that the proposed model discovers the causalities according to expert knowledge. For the prediction of the timeframes in the management of construction projects [82], researchers have proposed the application of AI instruments within the construction schedule [14]. In this study, an original optimization dispersion search algorithm was presented, which takes into account both the technological and the organizational constraints.

### C. Human Resources

Within the field of project management, human resource management is crucial since the projects depend on having the best possible human capital. One study [116] provided a new approach to the evaluation and classification of candidates during the recruitment process, which involves estimating their emotional intelligence using the data from social networks. Elsewhere, in [82], the focus was on efficient classification algorithms to predict employee performance

TABLE I. Main Studies of each Research Area

| Category | Investigation | Optimization Techniques Used |
|---|---|---|
| **Tenders** | | |
| **Predicting project award price** | [36] | (NN)+(CBR) |
| **Sales Prediction** | [148] | (SVM), (NN) |
| **Project** | | |
| **Project data analytics** | [28]<br>[7] | (EFNIM)<br>(Bootstrap) |
| **Project risk modeling, mitigation and management** | [72]<br>[34] | (BN), (BNCC)<br>(GA)+(SVM) |
| **Project mitigation and recovery plans** | [93] | (ANN)+(CBR) |
| **Project execution discovery and modeling** | [11]<br>[94] | (GA)+(CPM)<br>(GA)+(FL)+(NN)=(EFNIM) |
| **Real time predictive analytics** | [69]<br>[32] | (GA)<br>(EFHNN) |
| **Agile Project Management** | [44] | (CNN) |
| **Automated report generation** | [45] | (GA) |
| **Human Resources** | | |
| **Candidate identification and screening** | [116] | (DT), (SVM) and (BN). |
| **Performance management** | [82] | (DT) |
| **Retention management** | [78] | (DT) |
| **HR analytics** | [140] | (ANN) |
| **Information Technology** | | |
| **Cybersecurity prediction and analytics** | [149]<br>[133] | (ANN)+(BLN)+(SOM)<br>(ANN), (FL), (DT), (KNN), (SVM) |
| **Knowledge management** | [21] | (ANN)+(FL)+(GA) |
| **Design recognition library** | [114] | (GA) |
| **Innovation support and prioritization** | [141] | (ANN)+(FL)+(GA) |
| **Logistics** | | |
| **Automated Logistical Truck Services** | [3] | (RNN), (CNN) |
| **Object Detection and Classification Avoidance and Navigation** | [15] | (ACO), (AG) (ANN), (AS). (AIS)- (FNN) |

| Category | Investigation | Optimization Techniques Used |
|---|---|---|
| **Engineering & Design** | | |
| **Planning** | [6]<br>[105]<br>[118] | (ANN)<br>(GA)+(TS)<br>(GA) |
| **Stakeholder Management** | [33] | (EFSIM) |
| **Estimating** | [80]<br>[107] | (MA)<br>(ANN)+(FL) |
| **Design automation and optimization** | [134]<br>[164]<br>[9]<br>[83]<br>[115]<br>[131] | (ANN)+(GA)<br>(GA), (PSO), (SA), (AIS), (HS)<br>(ANN)<br>(ANN)<br>(ANN)+(GA)<br>(PNN) |
| **Generative design** | [110] | (ANN), (GA), (BN), (SVM), (HS) |
| **Continuous improvement** | [117] | (WOA) |
| **Evolving skills** | [25] | (wSVM)+(FL) +(fmGA) |
| **Operations** | | |
| **Back office/ automation/ Facilities management** | [157] | (MLR), (ANN), (SVM), (HS) |
| **Predictive maintenance** | [156]<br>[106] | (ANN)+(FL)+(GA)<br>(ANN)+(FL)+(GA)+(CBR) |
| **Operating project analytics** | [16] | (ANN)+(FL)+(GA) |
| **Autonomous systems** | [121] | (ANN)+(FL)+(GA)+(PSO) |
| **Supply Chain** | | |
| **Supply Chain** | [161] | (ANN) |
| **Construction** | | |
| **Construction management** | [76] | (ANN) |
| **Construction cost estimation** | [89]<br>[88]<br>[139] | (CBR)+(GA)<br>(ANN)+(CBR)+(MRA)<br>(MLP)+(GPA) |
| **Construction risk management** | [68] | (Bootstrap) |
| **Construction contract management** | [39] | (CBR) |
| **Construction safety** | [127] | (ANN) |
| **Project portfolio selection** | [1]<br>[138] | (CBR)+(FL)<br>(HNN)+(PSO) |
| **Onsite supervisory manpower/ Management** | [23]<br>[22] | (ANN)+(CBR)<br>(ANN)+(CBR) |

and on the mining that is commonly used in many areas and has been carried out by applying decision tree and classification algorithms for predicting employee performance.

### *D. Information Technology*

Information technology is a new area within project management but is one that is as important as all the other processes. A study was carried out [149] in relation to an implementation model for computer and network security purposes. Here, the aim was to use the model to combat malicious user activity. A smart hybrid system based on Bayesian learning networks and self-organizing maps was created and used to classify the networks and the host-based data collected within a local area network. Elsewhere, a study on cybersecurity and the optimization in smart "autonomous" buildings [124] explored the opportunities and challenges related to cybersecurity in Internet of Things (EIoT) environments in terms of the energy in smart buildings. Here, the proposed model can make decisions based on the data from neural networks that are designed with a circuit feedback loop with the ability to learn over time, which allows for learning from defined datasets and making smart decisions.

### *E. Engineering and Design*

AI methods have been used for the optimization of hybrid energy systems [164] and models (evolutionary diffuse SVM) for estimating

the construction costs. It is essential to monitor the project costs and to identify any potential problems.

### F. Operations

Operation and maintenance are also important aspects of industrial projects, and numerous studies show how AI affects future predictive maintenance. Here, one study [156] discusses the impact of AI on predictive maintenance, which is an important aspect of advanced production systems.

### G. Supply Chain

A two-stage methodology has been applied to an industrial survey dataset to investigate the relationships between key factors in a supply chain model [161]. The advantage of this model is that it frees the researcher from making subjective decisions during the analysis in terms of, for example, specifying the acceptable initial route models required for standard analysis.

### H. Logistics

Researchers have conducted a general analysis of the AI techniques applied throughout the world to address transportation issues, primarily in terms of traffic management, traffic safety, public transportation, and urban mobility [3]. Further studies on the management of warehouses using AI have also been conducted [15], while DHL also proposed an interesting approach in [62].

### I. Construction

Neural networks are regarded as a promising management tool that can enhance the current automation efforts in project management [76], the construction phase, and the engineering phase [63]. Studies on AI have also been carried out to identify the security risks in construction, with a focus on the management of the portfolio of projects using AI while taking into account the factors that generate risk in industrial projects and the historical records of the company [1].

## IV. Conclusions

The possibility of project success is a field of research in which researchers are working intensively. Here, the initial approaches were based on statistical models that have not responded to the needs of project management. In the field of AI, researchers have identified the algorithms and tools that can best deal with the various project variables and complex environments, with specific algorithms devised to address specific problems in the project. The main conclusions drawn from the reviewed works include that AI tools are more precise than traditional tools, while, at present, they remain somewhat complementary to the traditional approaches.

AI tools are highly useful to the project manager in terms of controlling and monitoring the project; however, many of the reviewed models involve weaknesses and limitations, which indicates that project managers should continue to use their experience when making evaluations according to the results. The trend of merging different AI tools continues to hold sway, wherein the strengths of one tool can compensate for the weaknesses of another. Indeed, this approach is returning the best results, and this is where the future lies. In this work, we studied the available AI techniques and the possible applications in the field of project management. In future work, a hybrid computational model that could fully ascertain the potential of AI in the field of project management will be proposed. The hope is that the management of autonomous projects will only require the partial supervision of a human project manager.

However, an autonomous project management system will also need to consider and fully control the project environment, including in terms of the status of the customers or the project stakeholders. Such a system can be used to apply AI algorithms for psychological and emotional analysis to evaluate both team performance and customer satisfaction. Looking to the future of 25 years from now, it is likely that there will exist an AI capable of managing the entire project, albeit with some form of human supervision.

The slow progress of AI in the field of project management is largely due to the lack of investment from private companies, which means progress is only been made in the universities and the public research organizations. In the future, AI will make all the decisions and will manage the resources in an optimal and timely manner, while the project manager will take the role of data scientist, working as part of a team with the AI to interpret the data and the decision making. Overall then, project managers will continue to play a crucial role when the AI is fully developed.

## Acknowledgments

I wish to thank my thesis tutors for the great help they provided. I also thank my family, especially my father, Francisco Gil Moreno, a successful businessman, without whom I could not have got where I am today, may he rest in peace.

## References

[1] H. R. Abbasianjahromi and H. Rajaie, "Application of fuzzy cbr and modm approaches in the project portfolio selection in construction companies," Iranian Journal of Science and Technology - Transactions of Civil Engineering, vol. 37, no. C1, pp. 143–155, 2013.

[2] A. Y. Abdelaziz, M. Z. Kamh, S. F. Mekhamer, and M. A. L. Badr, "A hybrid HNN-QP approach for dynamic economic dispatch problem," Electric Power Systems Research., vol. 78, no. 10, pp. 1784–1788, 2008, doi: 10.1016/j.epsr.2008.03.011.

[3] R. Abduljabbar, H. Dia, S. Liyanage, and S. A. Bagloee, "Applications of artificial intelligence in transport: An overview," Sustainability (Switzerland), vol. 11, no. 1. 2019, doi: 10.3390/su11010189.

[4] A. Abraham, "Adaptation of Fuzzy Inference System Using Neural Learning, in Fuzzy Systems Engineering: Theory and Practice, Studies in Fuzziness and Soft Computing," Studies in Fuzziness and Soft Computing, vol. 181, no. 3, 2005.

[5] B. Akay and D. Karaboga, "Artificial bee colony algorithm for large-scale problems and engineering design optimization," Journal of Intelligent Manufacturing, vol. 23, pp. 1001–1014, 2012, doi: 10.1007/s10845-010-0393-4.

[6] F. Amer and M. Golparvar-Fard, "Formalizing Construction Sequencing Knowledge and Mining Company-Specific Best Practices from Past Project Schedules," in Computing in Civil Engineering 2019: Visualization, Information Modeling, and Simulation - Selected Papers from the ASCE International Conference on Computing in Civil Engineering 2019, 2019, pp. 215–223, doi: 10.1061/9780784482421.028.

[7] L. Angelis and I. Stamelos, "A simulation tool for efficient analogy based cost estimation," Empirical Software Engineering, vol. 5, no. 1, pp. 35–68, 2000, doi: 10.1023/A:1009897800559.

[8] D. Anguita, A. Ghio, N. Greco, L. Oneto, and S. Ridella, "Model selection for support vector machines: Advantages and disadvantages of the Machine Learning Theory," in Proceedings of the International Joint Conference on Neural Networks, 2010, doi: 10.1109/IJCNN.2010.5596450.

[9] O. Arslan and O. Yetik, "ANN based optimization of supercritical ORC-Binary geothermal power plant: Simav case study," in Applied Thermal Engineering, 2011, vol. 31, no. 17–18, pp. 3922–3928, doi: 10.1016/j.applthermaleng.2011.07.041.

[10] L. Auria and R. A. Moro, "support vector machine as a Technique of Solvency analysis," DIW Berlin, vol. 811, 2008. doi: 10.2139/ssrn.1424949.

[11] R. F. Aziz, S. M. Hafez, and Y. R. Abuel-Magd, "Smart optimization for mega construction projects using artificial intelligence," Alexandria Eng. J., vol. 53, no. 3, pp. 591–606, 2014, doi: 10.1016/j.aej.2014.05.003.

[12] R. Bhattacharyya, P. Kumar, and S. Kar, "Fuzzy R&D portfolio selection of

interdependent projects," Computers and Mathematics with Applications, vol. 62, no. 10, pp. 3857–3870, 2011, doi: 10.1016/j.camwa.2011.09.036.T.

[13] Bhoskar, O. K. Kulkarni, N. K. Kulkarni, S. L. Patekar, G. M. Kakandikar, and V. M. Nandedkar, "Genetic Algorithm and its Applications to Mechanical Engineering: A Review," in Materials Today: Proceedings, 2015, vol. 2, no. 4–5, pp. 2624–2630, doi: 10.1016/j.matpr.2015.07.219.

[14] W. Boejko, Z. Hejducki, and M. Wodecki, "Applying metaheuristic strategies in construction projects management," Journal of Civil Engineering and Management, vol. 18, no. 5, pp. 621–630, Oct. 2012, doi: 10.3846/13923730.2012.719837.

[15] E. Bottani, R. Montanari, M. Rinaldi, and G. Vignali, "Intelligent algorithms for warehouse management," Intelligent Systems Reference Library, vol. 87, pp. 645–667, 2015, doi: 10.1007/978-3-319-17906-3_25.

[16] G. Braswell, "Artificial Intelligence Comes of Age in Oil and Gas," Journal of Petroleum Technology, vol. 65, no. 01, pp. 50–57, 2013, doi: 10.2118/0113-0050-jpt.

[17] S. T. Buckland, A. C. Davison, and D. V. Hinkley, "Bootstrap Methods and Their Application," Biometrics, vol. 52, no. 2, p. 795, 1998, doi: 10.2307/3109789.

[18] E. Carmona, "Tutorial sobre Maquinas de Vectores Soporte (SVM)." 2016. UNED, Consultada en http://www. ia. uned. es/~ ejcarmona/publicaciones/[2013-Carmona]%20SVM.pdf (fecha de consulta 01-07-2017).

[19] A. Castillo, J.M, Cortes, C, Gonzalez, J., & Benito, "Prospecting The Future with AI," Internacional Journal of Artificial Intelligence and Interactive Multimedia, vol. 1, no. 2, pp. 1–53, 2009.

[20] C. Chapell and V. Vapnik, "Model selection for Support Vector Machines," in Advances in Neural Information Processing Systems, 2000, pp. 230–236, doi:10.5555/3009657.3009690

[21] [1] S. S. Chaudhry, M. W. Varano, and L. Xu, "Systems research, genetic algorithms and information systems," Systems Research and Behavioral Science, 2000, doi: 10.1002/(sici)1099-1743(200003/04)17:2<149::aid-sres290>3.3.co;2-h.

[22] J. H. Chen and S. C. Hsu, "Hybrid ANN-CBR model for disputed change orders in construction projects," Automation in Construction, 2007, doi: 10.1016/j.autcon.2007.03.003.

[23] J. H. Chen, L. R. Yang, W. H. Chen, and C. K. Chang, "Case-based allocation of onsite supervisory manpower for construction projects," Construction Management and Economics, 2008, doi: 10.1080/01446190802014778.

[24] M. Y. Cheng, J. S. Chou, A. F. V. Roy, and Y. W. Wu, "High-performance Concrete Compressive Strength Prediction using Time-Weighted Evolutionary Fuzzy Support Vector Machines Inference Model," Automation in Construction, 2012, doi: 10.1016/j.autcon.2012.07.004..

[25] M. Y. Cheng, N. D. Hoang, A. F. V. Roy, and Y. W. Wu, "A novel time-depended evolutionary fuzzy SVM inference model for estimating construction project at completion," Engineering Applications of Artificial Intelligence, vol. 25, no. 4, pp. 744–752, 2012, doi: 10.1016/j.engappai.2011.09.022.

[26] M. Y. Cheng, L. C. Lien, H. C. Tsai, and P. H. Chen, "Artificial intelligence approaches to dynamic project success assessment taxonomic," Life Science Journal, vol. 9, pp. 5156–5163, 2012.

[27] M. Y. Cheng, H. C. Tsai, and W. S. Hsieh, "Web-based conceptual cost estimates for construction projects using Evolutionary Fuzzy Neural Inference Model," Automation in Construction, vol. 18, no. 2, pp. 164–172, 2009, doi: 10.1016/j.autcon.2008.07.001.

[28] M. Y. Cheng, H. C. Tsai, and C. L. Liu, "Artificial intelligence approaches to achieve strategic control over project cash flows," Automation in Construction, vol. 18, no. 4, pp. 386–393, 2009, doi: 10.1016/j.autcon.2008.10.005.

[29] M. Y. Cheng, H. C. Tsai, and E. Sudjono, "Evaluating subcontractor performance using evolutionary fuzzy hybrid neural network," International Journal of Project Management, vol. 29, no. 3, pp. 349–356, 2011, doi: 10.1016/j.ijproman.2010.03.005.

[30] M. Y. Cheng, H. C. Tsai, and E. Sudjono, "Evolutionary fuzzy hybrid neural network for conceptual cost estimates in construction projects," in 2009 26th International Symposium on Automation and Robotics in Construction, ISARC 2009, 2009, pp. 512–519, doi: 10.22260/isarc2009/0040.

[31] M. Y. Cheng, D. K. Wibowo, D. Prayogo, and A. F. V. Roy, "Predicting productivity loss caused by change orders using the evolutionary fuzzy support vector machine inference model," Journal of Civil Engineering and Management, vol. 21, no. 7, pp. 881–892, Oct. 2015, doi: 10.3846/13923730.2014.893922.

[32] T. M. Cheng and R. Z. Yan, "Integrating messy genetic algorithms and simulation to optimize resource utilization," Computer-Aided Civil and Infrastructure Engineering, vol. 24, no. 6, pp. 401–415, 2009, doi: 10.1111/j.1467-8667.2008.00588.x..

[33] J. S. Chou, M. Y. Cheng, and Y. W. Wu, "Improving classification accuracy of project dispute resolution using hybrid artificial intelligence and support vector machine models," Expert Systems with Applications, vol. 40, no. 6, pp. 2263–2274, 2013, doi: 10.1016/j.eswa.2012.10.036.

[34] J. S. Chou, M. Y. Cheng, Y. W. Wu, and A. D. Pham, "Optimizing parameters of support vector machine using fast messy genetic algorithm for dispute classification," Expert Syst. Appl., 2014, doi: 10.1016/j.eswa.2013.12.035.

[35] J. S. Chou, M. Y. Cheng, Y. W. Wu, and A. D. Pham, "Optimizing parameters of support vector machine using fast messy genetic algorithm for dispute classification," Expert Systems with Applications, vol. 41, no. 8, pp. 3955–3964, 2014, doi: 10.1016/j.eswa.2013.12.035.

[36] J. S. Chou and A. D. Pham, "Smart Artificial Firefly Colony Algorithm-Based Support Vector Regression for Enhanced Forecasting in Civil Engineering," Computer-Aided Civil and Infrastructure Engineering, vol. 30, no. 9, pp. 715–732, 2015, doi: 10.1111/mice.12121.

[37] J. S. Chou and A. D. Pham, "Smart Artificial Firefly Colony Algorithm-Based Support Vector Regression for Enhanced Forecasting in Civil Engineering," Comput. Civ. Infrastruct. Eng., 2015, doi: 10.1111/mice.12121.

[38] J. S. Chou, A. D. Pham, and H. Wang, "Bidding strategy to support decision-making by integrating fuzzy AHP and regression-based simulation," Automation in Construction, vol. 35, pp. 517–527, 2013, doi: 10.1016/j.autcon.2013.06.007.

[39] D. K. Chua and P. K. Loh, "CB-Contract: Case-Based Reasoning Approach to Construction Contract Strategy Formulation," Journal of Computing in Civil Engineering, vol. 20, no. 5, pp. 339–350, 2006, doi: 10.1061/(asce)0887-3801(2006)20:5(339).

[40] O. Cordón, "A historical review of evolutionary learning methods for Mamdani-type fuzzy rule-based systems: Designing interpretable genetic fuzzy systems," International Journal of Approximate Reasoning, vol. 52, no. 6. pp. 894–913, 2011, doi: 10.1016/j.ijar.2011.03.004.

[41] F. Costantino, G. Di Gravio, and F. Nonino, "Project selection in project portfolio management: An artificial neural network model based on critical success factors," International Journal of Project Management, vol. 33, no. 8, pp. 1744–1754, 2015, doi: 10.1016/j.ijproman.2015.07.003.

[42] E. Cox, Fuzzy Modeling and Genetic Algorithms for Data Mining and Exploration. 2005. Doi: 10.1016/B978-0-12-194275-5.X5000-2.

[43] A. L. A. Dalhoum and M. Al-Rawi, "High-Order Neural Networks are Equivalent to Ordinary Neural Networks," Modern Applied Science, vol. 13, no. 2, p. 228, Jan. 2019, doi: 10.5539/mas.v13n2p228.

[44] H. K. Dam, T. Tran, J. Grundy, A. Ghose, and Y. Kamei, "Towards effective AI-powered agile project management," in Proceedings - 2019 IEEE/ACM 41st International Conference on Software Engineering: New Ideas and Emerging Results, ICSE-NIER 2019, 2019, pp. 41–44, doi: 10.1109/ICSE-NIER.2019.00019.

[45] R. Day, J. Zydallis, G. Lamont, and R. Pachter, "Analysis of fine granularity and building block sizes in the parallel fast messy GA," in Proceedings of the 2002 Congress on Evolutionary Computation, CEC 2002, 2002, vol. 1, pp. 127–132, doi: 10.1109/CEC.2002.1006221.

[46] J. J. G. De la Rosa, A. A. Pérez, J. C. Palomares Salas, J. G. Ramiro Leo, and A. M. Muñoz, "A novel inference method for local wind conditions using genetic fuzzy systems," Renewable Energy, vol. 36, no. 6, pp. 1747–1753, 2011, doi: 10.1016/j.renene.2010.12.017.

[47] N. Dong, M. Fischer, D. Ge, and R. E. Levitt, "Automated look-ahead schedule generation and optimization for the finishing phase of complex construction projects," CIFE Technical Report, Stanford University, no. June, p. 1 online resource., 2012.

[48] S. Ebrahimnejad, S. M. Mousavi, and H. Seyrafianpour, "Risk identification and assessment for build-operate-transfer projects: A fuzzy multi attribute decision making model," Expert Systems with Applications, vol. 37, no. 1, pp. 575–586, Jan. 2010, doi: 10.1016/j.eswa.2009.05.037.

[49] T. Efendigil, S. Önüt, and C. Kahraman, "A decision support system for demand forecasting with artificial neural networks and neuro-fuzzy

models: A comparative analysis," Expert Systems with Applications, vol. 36, no. 3 PART 2, pp. 6697–6707, 2009, doi: 10.1016/j.eswa.2008.08.058.

[50] B. Efron and R. Tibshirani, "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy," Statistical Science, vol. 1, no. 1, pp. 54–75, 1986, doi: 10.1214/ss/1177013815.

[51] C. Egbu and S. Suresh, "Knowledge Mapping Techniques Within The Construction Industry:An Exploratory Study," CIB W102-Information Knowl. Manag. Build., pp. 48–57, 2008.

[52] C. Egbu and S. Suresh, "Knowledge Mapping Techniques Within The Construction Industry:An Exploratory Study," CIB W102-Information and knowledge management in Buildings, pp. 48–57, 2008.

[53] A. O. Esogbue and J. A. Murrell, "Fuzzy adaptive controller using reinforcement learning neural networks," in 1993 IEEE International Conference on Fuzzy Systems, 1993, pp. 178–183, doi: 10.1109/fuzzy.1993.327494.

[54] W. H. Estler H.C, "Heuristic Search-Based Planning for Graph Transformation Systems," ICAPS Workshop on Knowledge Engineering for Planning and Scheduling, p. 54, 2011.

[55] A. M. Fahim, A. M. Salem, F. A. Torkey, and M. A. Ramadan, "Efficient enhanced k-means clustering algorithm," Journal of Zhejiang University: Science, vol. 7, no. 10, pp. 1626–1633, 2006, doi: 10.1631/jzus.2006.A1626.

[56] E. Faliagka et al., "On-line consistent ranking on e-recruitment: Seeking the truth behind a well-formed CV," Artificial Intelligence Review, vol. 42, no. 3, pp. 515–528, 2014, doi: 10.1007/s10462-013-9414-y.

[57] M. Fazzolari, R. Alcala, Y. Nojima, H. Ishibuchi, and F. Herrera, "A review of the application of multiobjective evolutionary fuzzy systems: Current status and further directions," IEEE Transactions on Fuzzy Systems, vol. 21, no. 1, pp. 45–65, 2013, doi: 10.1109/TFUZZ.2012.2201338.

[58] C. M. Fonseca and P. J. Fleming, "Genetic Algorithms for Multiobjective Optimization: Formulation, Discussion and Generalization," Icga, vol. 93, no. July, pp. 416–423, 1993, doi: citeulike-article-id:2361311.

[59] A. Fukunaga, G. Rabideau, S. Chien, and D. Yan, "Towards an application framework for automated planning and scheduling," in IEEE Aerospace Applications Conference Proceedings, 1997, vol. 1, pp. 375–386, doi: 10.1109/aero.1997.574426.

[60] T. Ganesan, P. Vasant, and I. Elamvazuthi, "Hopfield neural networks approach for design optimization of hybrid power systems with multiple renewable energy sources in a fuzzy environment," Journal of Intelligent and Fuzzy Systems, vol. 26, no. 5, pp. 2143–2154, 2014, doi: 10.3233/IFS-130889.

[61] B. S. Garcia, R., Perez, I., Villavicencio, N., Piñero, P., "Experiences by using genetic algorithms in project scheduling," Revista Cubana de Ciencias Informáticas, vol. 10, pp. 71–86, 2016.

[62] L. Kota, "Artificial Intelligence in Logistics," Advanced Logistic Systems - Theory and Practice, vol. 12, no. 1, pp. 47–60, 2019, doi: 10.32971/als.2019.004.

[63] Y. M. Goh and D. K. H. Chua, "Case-Based Reasoning Approach to Construction Safety Hazard Identification: Adaptation and Utilization," Journal of Construction Engineering and Management, vol. 136, no. 2, pp. 170–178, 2010, doi: 10.1061/(asce)co.1943-7862.0000116.

[64] D. Goldberg and K. Deb, "Rapid, accurate optimization of difficult problems using messy genetic algorithms," proceedings of the fifth international conference on genetic algorithms, pp. 56–64, 1993.

[65] D. GOLDBERG, "Messy Genetic Algorithms: Motivation, Analysis, and First Results," Complex Systems, vol. 3, no. 5, pp. 493–530, 1989.

[66] F. Hartman and R. A. Ashrafi, "Project Management in the Information Systems and Information Technologies Industries," Project Management Journal, 2002, doi: 10.1177/875697280203300303.

[67] H. Hashemi, S. M. Mousavi, and S. M. H. Mojtahedi, "Bootstrap technique for risk analysis with interval numbers in bridge construction projects," J. Constr. Eng. Manag., 2011, doi: 10.1061/(ASCE)CO.1943-7862.0000344.

[68] H. Hashemi, S. M. Mousavi, R. Tavakkoli-Moghaddam, and Y. Gholipour, "Compromise Ranking Approach with Bootstrap Confidence Intervals for Risk Assessment in Port Management Projects," Journal of Management in Engineering, vol. 29, no. 4, pp. 334–344, 2013, doi: 10.1061/(asce)me.1943-5479.0000167.

[69] T. Hegazy, "Optimization of construction time - Cost trade-off analysis using genetic algorithms," Canadian Journal of Civil Engineering, vol. 26, no. 6, pp. 685–697, 1999, doi: 10.1139/l99-031.

[70] D. Heiss-Czedik, "An Introduction to Genetic Algorithms.," Artificial Life, vol. 3, no. 1, pp. 63–65, 1997, doi: 10.1162/artl.1997.3.63.

[71] F. Herrera, "Genetic fuzzy systems: Taxonomy, current research trends and prospects," Evolutionary Intelligence, vol. 1, no. 1, pp. 27–46, 2008, doi: 10.1007/s12065-007-0001-5.

[72] X. Hu, B. Xia, M. Skitmore, and Q. Chen, "The application of case-based reasoning in construction management research: An overview," Automation in Construction, vol. 72. pp. 65–74, 2016, doi: 10.1016/j.autcon.2016.08.023.

[73] Y. Hu, X. Zhang, E. W. T. Ngai, R. Cai, and M. Liu, "Software project risk analysis using Bayesian networks with causality constraints," Decision Support Systems, vol. 56, no. 1, pp. 439–449, Dec. 2013, doi: 10.1016/j.dss.2012.11.001.

[74] S. V. Ioannou, A. T. Raouzaiou, V. A. Tzouvaras, T. P. Mailis, K. C. Karpouzis, and S. D. Kollias, "Emotion recognition through facial expression analysis based on a neurofuzzy network," Neural Networks, vol. 18, no. 4, pp. 423–435, 2005, doi: 10.1016/j.neunet.2005.03.004.

[75] H. Ishibuchi, "Genetic fuzzy systems: evolutionary tuning and learning of fuzzy knowledge bases," Fuzzy Sets and Systems, vol. 141, no. 1, pp. 161–162, 2004, doi: 10.1016/s0165-0114(03)00262-8.

[76] M. Jain and K. K. Pathak, "Applications of artificial neural network in construction engineering and management - A review," International Journal of Engineering Technology, Management and Applied Sciences, vol. 2, no. 3, pp. 134–142, 2014.

[77] H. R. Jantan, A. A. Hamdan, and Z. Othman, "Human Talent Forecasting using Data Mining Classification Techniques," International Journal of Technology Diffusion, vol. 1, no. 4, pp. 29–41, 2011, doi: 10.4018/jtd.2010100103.

[78] H. Jantan, "Human Talent Prediction in HRM using C4 . 5 Classification Algorithm," International Journal on Computer Science and Engineering, vol. 02, no. 08, pp. 2526–2534, 2010.

[79] C. D. Jeffries, "Tracking, code recognition, and memory management with high-order neural networks",Applications and Science of Artificial Neural Networks, 1995, vol. 2492, pp. 964–973, doi: 10.1117/12.205207.

[80] L. Jin, C. Zhang, X. Shao, and G. Tian, "Mathematical modeling and a memetic algorithm for the integration of process planning and scheduling considering uncertain processing times," Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture, vol. 230, no. 7, pp. 1272–1283, 2016, doi: 10.1177/0954405415625916.

[81] V. Kachitvichyanukul, "Comparison of Three Evolutionary Algorithms: GA, PSO, and DE," Industrial Engineering and Management Systems, vol. 11, no. 3, pp. 215–223, 2012, doi: 10.7232/iems.2012.11.3.215.

[82] V. Kalaivani and M. M. Elamparithi, "An Efficient Classification Algorithms for Employee Performance Prediction," International Journal of Research in Advent Technology, vol. 2, no. 9, pp. 27–32, 2014.

[83] S. A. Kalogirou, "Artificial neural networks and genetic algorithms for the optimisation of solar thermal systems," in Artificial Intelligence in Energy and Renewable Energy Systems, Nova Science Publishers, Inc., 2006, pp. 131–162.

[84] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithms: Analysis and implementation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 881–892, 2002, doi: 10.1109/TPAMI.2002.1017616.

[85] S. Kar, S. Das, and P. K. Ghosh, "Applications of neuro fuzzy systems: A brief review and future outline," Applied Soft Computing Journal, vol. 15. pp. 243–259, 2014, doi: 10.1016/j.asoc.2013.10.014.

[86] H. Kargupta, "SEARCH, polynomial complexity, and the fast messy genetic algorithm," Urbana, vol. 51, no. 95008, p. 188, 1996.

[87] N. A. Kartam, R. E. Levitt, and D. E. Wilkins, "Extending Artificial Intelligence Techniques for Hierarchical Planning," Journal of Computing in Civil Engineering, vol. 5, no. 4, pp. 464–477, 1991, doi: 10.1061/(asce)0887-3801(1991)5:4(464).

[88] G. H. Kim, S. H. An, and K. I. Kang, "Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning," Build. Environ., 2004, doi: 10.1016/j.buildenv.2004.02.013.

[89] G. H. Kim, S. H. An, and K. I. Kang, "Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning," Building and Environment, vol. 39, no. 10, pp. 1235–1242, 2004, doi: 10.1016/j.buildenv.2004.02.013.

[90] O. Kisi, "Modeling solar radiation of Mediterranean region in Turkey by using fuzzy genetic approach," Energy, vol. 64, pp. 429–436, 2014, doi: 10.1016/j.energy.2013.10.009.

[91] M. Klumpp, "Automation and artificial intelligence in business logistics systems: human reactions and collaboration requirements," International Journal of Logistics Research and Applications, vol. 21, no. 3, pp. 224–242, 2018, doi: 10.1080/13675567.2017.1384451.

[92] C. H. Ko and M. Y. Cheng, "Hybrid use of AI techniques in developing construction management tools," Automation in Construction, vol. 12, no. 3, pp. 271–281, May 2003, doi: 10.1016/S0926-5805(02)00091-2.

[93] C.-H. Ko and M.-Y. Cheng, "Dynamic Prediction of Project Success Using Artificial Intelligence," Journal of Construction Engineering and Management, vol. 133, no. 4, pp. 316–324, 2007, doi: 10.1061/(asce)0733-9364(2007)133:4(316).

[94] C. H. Ko, M. Y. Cheng, and T. K. Wu, "Evaluating sub-contractors performance using EFNIM," Automation in Construction, vol. 16, no. 4, pp. 525–530, 2007, doi: 10.1016/j.autcon.2006.09.005.

[95] K. A. H. Kobbacy, "Application of Artificial Intelligence in maintenance modelling and management," in IFAC Proceedings Volumes (IFAC-PapersOnline), 2012, doi: 10.3182/20121122-2-ES-4026.00046.

[96] K. A. H. Kobbacy, "Application of Artificial Intelligence in maintenance modelling and management," in IFAC Proceedings Volumes (IFAC-PapersOnline), 2012, vol. 45, no. 31, pp. 54–59, doi: 10.3182/20121122-2-ES-4026.00046.

[97] M. Kumar, M. Husain, N. Upreti, and D. Gupta, "Genetic Algorithm: Review and Application," SSRN Electronic Journal, vol. 2, no. 2, pp. 451–454, 2020, doi: 10.2139/ssrn.3529843.

[98] H. K. Kwan, "High-order feedbackward neural networks," 1991, pp. 49–51, doi: 10.1109/ciccas.1991.184277.

[99] H. Kwasnicka and M. Przewozniczek, "Multi population pattern searching algorithm: A new evolutionary method based on the idea of messy genetic algorithm," IEEE Trans. Evol. Comput., 2011, doi: 10.1109/TEVC.2010.2102038.

[100] M. Lahmann, "AI will transform project management. Are you ready?," Pwc Switzerland, 2018. [Online]. Available: https://www.pwc.ch/en/insights/risk/transformation-assurance-ai-will-transform-project-management-are-you-ready.html.

[101] M. Laumanns and J. Ocenasek, "Bayesian optimization algorithms for multi-objective optimization," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2002, vol. 2439, pp. 298–307, doi: 10.1007/3-540-45712-7_29.

[102] G. G. Lendaris, K. Mathia, and R. Saeks, "Linear Hopfield networks and constrained optimization," IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 29, no. 1, pp. 114–118, 1999, doi: 10.1109/3477.740171.

[103] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," IEEE Circuits and Systems Magazine, vol. 9, no. 3, pp. 32–50, 2009, doi: 10.1109/MCAS.2009.933854.

[104] M. Y. Leyva Vázquez, K. Pérez Teruel, A. Febles Estrada, and J. Gulín González, "Mapas cognitivos difusos para la selección de proyectos de tecnologías de la información," Contaduría y Administración, vol. 58, no. 4, pp. 95–117, 2013, doi: 10.1016/s0186-1042(13)71235-x.

[105] X. Li, L. Gao, and W. Li, "Application of game theory based hybrid algorithm for multi-objective integrated process planning and scheduling," Expert Systems with Applications, vol. 39, no. 1, pp. 288–297, 2012, doi: 10.1016/j.eswa.2011.07.019.

[106] Z. Li, Y. Wang, and K. S. Wang, "Intelligent predictive maintenance for fault diagnosis and prognosis in machine centers: Industry 4.0 scenario," Advances in Manufacturing, vol. 5, no. 4, pp. 377–387, 2017, doi: 10.1007/s40436-017-0203-8.

[107] R. . Lozada, "Redes neuronales y logica difusa aplicado a un sistema climatológico," Universidad nacional de San Agustin, 2017.

[108] P. C. K. Luk, K. C. Low, and A. Sayiah, "GA-based fuzzy logic control of a solar power plant using distributed collector fields," Renewable Energy, vol. 16, no. 1–4, pp. 765–768, 1999, doi: 10.1016/s0960-1481(98)00275-4.

[109] M. M. Mijwil, "Artificial Neural Networks Advantages and Disadvantages," Linkedin, no. March, 2018.

[110] V. Machairas, A. Tsangrassoulis, and K. Axarli, "Algorithms for optimization of building design: A review," Renewable and Sustainable Energy Reviews, vol. 31. pp. 101–112, 2014, doi: 10.1016/j.rser.2013.11.036.

[111] J. G. MacKinnon, "Bootstrap methods in econometrics," in Economic Record, 2006, vol. 82, no. SPEC. ISS. 1, doi: 10.1111/j.1475-4932.2006.00328.x.

[112] D. Magaña Martínez and J. C. Fernandez-Rodriguez, "Artificial Intelligence Applied to Project Success: A Literature Review," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 3, no. 5, p. 77, 2015, doi: 10.9781/ijimai.2015.3510.

[113] R. Mahajan and G. Kaur, "Neural Networks using Genetic Algorithms," International Journal of Computer Applications, vol. 77, no. 14, pp. 6–11, Sep. 2013, doi: 10.5120/13549-1153.

[114] M. G. Marchetta and R. Q. Forradellas, "An artificial intelligence planning approach to manufacturing feature recognition," CAD Computer Aided Design, vol. 42, no. 3, pp. 248–256, 2010, doi: 10.1016/j.cad.2009.11.007.

[115] A. Mellit, S. a Kalogirou, and M. Drif, "Application of neural networks and genetic algorithms for sizing of photovoltaic systems," Renewable Energy, vol. 35, no. 12, pp. 2881–2893, 2010, doi: 10.1016/j.renene.2010.04.017.

[116] V. M. Menon and H. A. Rahulnath, "A novel approach to evaluate and rank candidates in a recruitment process by estimating emotional intelligence through social media data," in 2016 International Conference on Next Generation Intelligent Systems, ICNGIS 2016, 2017, doi: 10.1109/ICNGIS.2016.7854061.

[117] S. Mirjalili and A. Lewis, "The Whale Optimization Algorithm," Advances in Engineering Software, vol. 95, pp. 51–67, 2016, doi: 10.1016/j.advengsoft.2016.01.008.

[118] J. R. Montoya-Torres, E. Gutierrez-Franco, and C. Pirachicán-Mayorga, "Project scheduling with limited resources using a genetic algorithm," International Journal of Project Management, vol. 28, no. 6, pp. 619–628, 2010, doi: 10.1016/j.ijproman.2009.10.003.

[119] A. Moreno and T. Redondo, "Text Analytics: the convergence of Big Data and Artificial Intelligence," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 3, no. 6, p. 57, 2016, doi: 10.9781/ijimai.2016.369.

[120] O. Moselhi, T. Hegazy, and P. Fazio, "Neural Networks as Tools in Construction," Journal of Construction Engineering and Management, vol. 117, no. 4, pp. 606–625, 1991, doi: 10.1061/(asce)0733-9364(1991)117:4(606).

[121] G. P. Moustris, S. C. Hiridis, K. M. Deliparaschos, and K. M. Konstantinidis, "Evolution of autonomous and semi-autonomous robotic surgical systems: A review of the literature," International Journal of Medical Robotics and Computer Assisted Surgery, vol. 7, no. 4. pp. 375–392, 2011, doi: 10.1002/rcs.408.

[122] K. N. Mutter, Z. M. Jafri, and A. A. Aziz, "Hopfield Neural Network (HNN) Improvement for color image recognition using multi-bitplane and multi-connect architecture," in Computer Graphics, Imaging and Visualisation: New Advances, CGIV 2007, 2007, pp. 403–407, doi: 10.1109/CGIV.2007.24.

[123] M. Mylrea and N. G. Gourisetti, "Cybersecurity and optimization in smart 'autonomous' buildings," in Autonomy and Artificial Intelligence: A Threat or Savior?, 2017, pp. 263–294.

[124] G. Napoles, "Algoritmo para mejorar la convergencia en Mapas Cognitivos Difusos Sigmoidales," Universidad Central "Marta Abreu" de Las Villas Facultad de Matemática, Física y Computación., 2014.

[125] D. Nau, S. Gupta, and W. Regli, "Manufacturing-Operation Planning Versus AI Planning," in AAAI Spring Symposium on Integrated Planning Applications, 1995.

[126] M. Obayashi, T. Nishida, T. Kuremoto, K. Kobayashi, and L. B. Feng, "A reinforcement learning system embedded agent with neural network-based multi-valued pattern memory structure," in ICCAS 2010 - International Conference on Control, Automation and Systems, 2010, pp. 176–181, doi: 10.1109/iccas.2010.5669888.

[127] D. A. Patel and K. N. Jha, "Neural Network Model for the Prediction of Safe Work Behavior in Construction Projects," Journal of Construction Engineering and Management, vol. 141, no. 1, p. 04014066, 2015, doi: 10.1061/(asce)co.1943-7862.0000922.

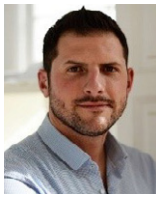
[128] M. Pelikan, D. E. Goldberg, and E. Cantú-Paz, "BOA : The Bayesian Optimization Algorithm 1 Introduction," Proceedings of the genetic and evolutionary computation conference GECCO-99, vol. 1, pp. 525–532, 1999.

[129] T. Pfeufer and M. Ayoubi, "Application of a hybrid neuro-fuzzy system to the fault diagnosis of an automotive electromechanical actuator," Fuzzy

Sets and Systems, vol. 89, no. 3, pp. 351–360, 1997, doi: 10.1016/S0165-0114(97)00022-5.
[130] PMI, PMBOK Guide | Project Management Institute. 2017.
[131] S. M. Pourkiaei, M. H. Ahmadi, and S. M. Hasheminejad, "Modeling and experimental verification of a 25W fabricated PEM fuel cell by parametric and GMDH-type neural network," Mech. Ind., 2016, doi: 10.1051/meca/2015050.
[132] R. Prieto, "Impacts of Artificial Intelligence on Management of Large Complex Projects," PM World Journal, vol. 8, no. 5, pp. 1–20, 2019.
[133] A. Rehman and T. Saba, "Evaluation of artificial intelligent techniques to secure information in enterprises," Artificial Intelligence Review, vol. 42, no. 4, pp. 1029–1044, 2014, doi: 10.1007/s10462-012-9372-9.
[134] C. Renzi, F. Leali, M. Cavazzuti, and A. O. Andrisano, "A review on artificial intelligence applications to the optimal design of dedicated and reconfigurable manufacturing systems," International Journal of Advanced Manufacturing Technology, vol. 72, no. 1–4, pp. 403–418, 2014, doi: 10.1007/s00170-014-5674-1.
[135] L. Ricalde, B. Cruz, and E. Sánchez, "Control neuronal recurrente de alto orden para turbinas de viento con generador síncrono de imán permanente," Computación y Sistemas, vol. 14, no. 2, pp. 133–143, 2010.
[136] L. Rodriguez-Repiso, R. Setchi, and J. L. Salmeron, "Modelling IT projects success with Fuzzy Cognitive Maps," Expert Systems with Applications, vol. 32, no. 2, pp. 543–559, 2007, doi: 10.1016/j.eswa.2006.01.032.
[137] L. Rutkowski, K. Cpałka, R. Nowicki, A. Pokropińska, and R. Scherer, "Neuro-fuzzy systems," in Computational Complexity: Theory, Techniques, and Applications, vol. 9781461418, 2012, pp. 2069–2081.
[138] A. N. Sadigh, H. Mokhtari, M. Iranpoor, and S. M. T. Fatemi Ghomi, "Cardinality constrained portfolio optimization using a hybrid approach based on particle swarm optimization and hopfield neural network," Advanced Science Letters, vol. 17, no. 1, pp. 11–20, 2012, doi: 10.1166/asl.2012.3666.
[139] E. L. O. Savaş Bayram, Mehmet Emin Öcal, "Analysis of Cost and Schedule Variances in Construction Works with Artificial Intelligence Approaches: The Case of Turkey," in International Students' Conference of Civil Engineering, ISCCE, 2012, pp. 10–11.
[140] T. Schmitt, P. Caillou, and M. Sebag, "Matching Jobs and Resumes: a Deep Collaborative Filtering Task," 2018, vol. 41, pp. 124–109, doi: 10.29007/17rz.
[141] S. G. Selivanov, S. N. Poezjalova, and O. A. Gavrilova, "The Use of Artificial Intelligence Methods of Technological Preparation of Engine-Building Production," American Journal of Industrial Engineering, vol. 2, no. 1, pp. 10–14, 2014, doi: 10.12691/AJIE-2-1-3.R.
[142] R. Sonmez, "Parametric Range Estimating of Building Costs Using Regression Models and Bootstrap," Journal of Construction Engineering and Management, vol. 134, no. 12, pp. 1011–1016, 2008, doi: 10.1061/(asce)0733-9364(2008)134:12(1011).
[143] C. D. Stylios and P. P. Groumpos, "Modeling Complex Systems Using Fuzzy Cognitive Maps," IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans., vol. 34, no. 1, pp. 155–162, 2004, doi: 10.1109/TSMCA.2003.818878.
[144] L. Suganthi, S. Iniyan, and A. A. Samuel, "Applications of fuzzy logic in renewable energy systems - A review," Renewable and Sustainable Energy Reviews, vol. 48. pp. 585–607, 2015, doi: 10.1016/j.rser.2015.04.037.
[145] M. Sugeno and T. Yasukawa, "A Fuzzy-Logic-Based Approach to Qualitative Modeling," IEEE Transactions on Fuzzy Systems, vol. 1, no. 1, pp. 7–31, 1993, doi: 10.1109/TFUZZ.1993.390281.
[146] H. K. Sulehria and Y. Zhang, "Hopfield Neural Networks — A Survey," Engineering, pp. 125–130, 2007.
[147] R. S. Sutton and A. G. Barto, "Reinforcement Learning: An Introduction," IEEE Transactions on Neural Networks, vol. 9, no. 5, pp. 1054–1054, Sep. 1998, doi: 10.1109/tnn.1998.712192.
[148] N. Syam and A. Sharma, "Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and practice," Ind. Mark. Manag., vol. 69, pp. 135–146, 2018, doi: 10.1016/j.indmarman.2017.12.019.
[149] J. L. Thames, R. Abler, and A. Saad, "Hybrid intelligent systems for network security," in Proceedings of the Annual Southeast Conference, 2006, vol. 2006, pp. 286–289, doi: 10.1145/1185448.1185513.
[150] M. K. Tiwari and C. Chatterjee, "Development of an accurate and reliable hourly flood forecasting model using wavelet-bootstrap-ANN (WBANN) hybrid approach," Journal of Hydrology, vol. 394, no. 3–4, pp. 458–470, 2010, doi: 10.1016/j.jhydrol.2010.10.001.
[151] M. K. Tiwari and C. Chatterjee, "Uncertainty assessment and ensemble flood forecasting using bootstrap based artificial neural networks (BANNs)," Journal of Hydrology, vol. 382, no. 1–4, pp. 20–33, 2010, doi: 10.1016/j.jhydrol.2009.12.013.
[152] A. Tsakonas and B. Gabrys, "A fuzzy evolutionary framework for combining ensembles," Applied Soft Computing Journal, vol. 13, no. 4, pp. 1800–1812, 2013, doi: 10.1016/j.asoc.2012.12.027.
[153] TutorialPoint.com, "Artificial Intelligence - Fuzzy Logic Systems." [Online]. Available: https://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_fuzzy_logic_systems.htm.
[154] H. Umit, "Digital business strategies in blockchain ecosystems : transformational design and future of global business (E-book)," Contributions to Management Science, pp. 569–599, 2000.
[155] N. T. T. Vu, N. P. Tran, and N. H. Nguyen, "Adaptive neuro-fuzzy inference system based path planning for excavator arm," Journal of Robotics, vol. 2018, 2018, doi: 10.1155/2018/2571243.
[156] K. Wang and Y. Wang, "How AI Affects the Future Predictive Maintenance: A Primer of Deep Learning," in Lecture Notes in Electrical Engineering, 2018, doi: 10.1007/978-981-10-5768-7_1.
[157] Z. Wang and R. S. Srinivasan, "A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models," Renewable and Sustainable Energy Reviews. 2017, doi: 10.1016/j.rser.2016.10.079.
[158] M. Wauters and M. Vanhoucke, "Support Vector Machine Regression for project control forecasting," Automation in Construction, vol. 47, pp. 92–106, 2014, doi: 10.1016/j.autcon.2014.07.014.
[159] R. Wehrens, H. Putter, and L. M. C. Buydens, "The bootstrap: A tutorial," Chemometrics and Intelligent Laboratory Systems, vol. 54, no. 1, pp. 35–52, 2000, doi: 10.1016/S0169-7439(00)00102-7.
[160] G. C. Whitley, D., Beveridge, J.R, GuerraSalcedo C., "Messy Genetic Algorithms for Subset Feature Selection," Department of Computer Science, pp. 568--575, 1997.
[161] T. C. Wong, K. M. Y. Law, H. K. Yau, and S. C. Ngan, "Analyzing supply chain operation models with the PC-algorithm and the neural network," Expert Systems with Applications, vol. 38, no. 6, pp. 7526–7534, 2011, doi: 10.1016/j.eswa.2010.12.115.
[162] J. A. Yacim and D. G. B. Boshoff, "Impact of artificial neural networks training algorithms on accurate prediction of property values," Journal of Real Estate Research, 2018.
[163] L. A. Zadeh, "Fuzzy sets," Information and Control, vol. 8, no. 3, pp. 338–353, 1965, doi: 10.1016/S0019-9958(65)90241-X.
[164] S. M. Zahraee, M. Khalaji Assadi, and R. Saidur, "Application of Artificial Intelligence Methods for Hybrid Energy System Optimization," Renewable and Sustainable Energy Reviews, vol. 66. pp. 617–630, 2016, doi: 10.1016/j.rser.2016.08.028.
[165] G. Zhang, B. Eddy Patuwo, and M. Y. Hu, "Forecasting with artificial neural networks: The state of the art," International Journal of Forecasting, vol. 14, no. 1, pp. 35–62, 1998, doi: 10.1016/S0169-2070(97)00044-7.
[166] J. Zhang, "Developing robust non-linear models through bootstrap aggregated neural networks," Neurocomputing, vol. 25, no. 1–3, pp. 93–113, 1999, doi: 10.1016/S0925-2312(99)00054-5.

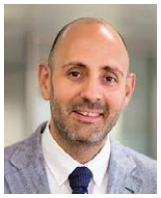
Jesús Gil Ruiz

Jesús Gil Ruiz is a PhD Candidate in computer science and is also an industrial engineer, civil engineer, and industrial organization engineer. He received an Msc Executive MBA, an Msc in Financial Management and Cost Control, and an Msc In Project, Construction, and Maintenance of Infrastructures and Facilities of Rail Lines from the University of Barcelona. Artificial Intelligence Program Applied to Strategic Management by MIT Management Executive Education and Business Analytics Program by Wharton Executive Education (University of Pennsylvania). He is a project manager in the renewable energy and oil and gas sectors, working in prestigious, internationally renowned companies such as TSK, Técnicas Reunidas, ABENGOA, and SENER Engineering and Systems. He has also participated in projects of great international importance, including BenBan Solar, the world's largest solar plant as of 2019 (1,800 MW), and Cauchari Solar (330MW), one of the largest photovoltaic plants built in Latin America in 2018. He is also an assistant professor at the University of La Rioja, where he works in the field of mathematics. He is also a trainer, lecturer, and speaker at Inicitivas Empresariales (an international company in Barcelona), working in the areas of project management, high-speed rail engineering and renewable energies.



Javier Martínez Torres

Dr. Javier Martínez Torres is a Mathematician and Engineering PhD from the University of Vigo. He is currently an Assistant Professor at the University of Vigo and has participated in more than 20 research projects as principal investigator. He has published more than 50 papers in JCR indexed journals and participate in more than 25 international conferences.



Rubén González Crespo

Dr. Rubén González Crespo has a PhD in Computer Science Engineering. Currently he is Vice Chancellor of Academic Affairs and Faculty from UNIR and Global Director of Engineering Schools from PROEDUCA Group. He is advisory board member for the Ministry of Education at Colombia and evaluator from the National Agency for Quality Evaluation and Accreditation of Spain (ANECA). He is member from different committees at ISO Organization. Finally, He has published more than 200 papers in indexed journals and congresses.

# An Effective Tool for the Experts' Recommendation Based on PROMETHEE II and Negotiation: Application to the Industrial Maintenance

Nawal Sad Houari[1]*, Noria Taghezout[2]

[1] Laboratoire d'Informatique Oran (LIO), Département du Vivant et de l'Environnement, Faculté des Sciences de la Nature et de la Vie, Université des Sciences et de la Technologie d'Oran Mohamed Boudiaf, USTO-MB, BP 1505, El M'naouer, 31000 Oran (Algeria)
[2] Laboratoire d'Informatique Oran (LIO), Université d'ORAN1 Ahmed Ben Bella, BP 1524 EL Mnaouer Oran (Algeria)

Received 2 August 2020 | Accepted 11 January 2021 | Published 17 January 2021



## Abstract

In this article, we propose an expert recommendation tool that relies on the skills of experts and their interventions in collaboration. This tool provides us with a list of the most appropriate (effective) experts to solve business problems in the field of industrial maintenance. The proposed system recommends experts using an unsupervised classification algorithm that takes into account the competences of the experts, their preferences and the stored information in previous collaborative sessions. We have tested the performance of the system with K-means and C-means algorithms. To fix the inconsistencies detected in business rules, the PROMETHEE II multi-criteria decision support method is integrated into the extended CNP negotiation protocol in order to classify the experts from best to worst. The study is supported by the well known petroleum company in Algeria namely SONATRACH where the experimentations are operated on maintenance domain. Experiments results show the effectiveness of our approach, obtaining a recall of 86%, precision of 92% and F-measure of 89%. Also, the proposed approach offers very high results and improvement, in terms of response time (154.28 ms), space memory (9843912 bytes) and negotiation rounds.

## Keywords

Business Rules, Expert's Skills, Fuzzy C-means, K-means, PROMETHEE II, Recommender Tool, SONATRACH, Unsupervised Classification.

DOI: 10.9781/ijimai.2021.01.002

## I. Introduction

AMONG the factors of company success, whatever its extension, is the existence of human capital of quality. Indeed, human is a more important resource than money. Companies that are not able to obtain and maintain a competent human capital, whatever the market case is, cannot evolve in an environment as shifting as the one where business is currently taking place.

This paper focuses on the evolution of a company by taking into account the experts business knowledge, this knowledge is a major capital for companies, the loss of this knowledge or its misuse potentially leads to the failure of the company.

The knowledge can be represented in several forms, for example in the form of business rules. Business rules can model a business decision. They capitalize a company's knowledge and translate its strategy by describing the actions to be taken for a given process. They are usually written in a controlled natural language. A business rule is a high-level description of how to control and / or make a decision, using specific concepts to a company or an organization. Thus, business rules describe what an expert must do to make a decision [1].

These business rules are usually housed in traditional computer programs, business processes and reference documents, and especially in the minds of business experts (i.e. from expertise). In our case, we were interested in the capitalization of the knowledge possessed by the business experts via the creation of a business rules management system.

In order to properly identify our problematic, we will present the real reasons for the need of a skills-based recommendation tool and performance monitoring to manage the business rules in an enterprise. The main problematic treated in this paper is to provide fast solutions to inconsistent business rules where some business experts who work in collaboration are unable to adhere to an idea or a suggestion related to the business rule. In fact, setting up a dynamic negotiation protocol is not an easy task, the fact of calling all the enterprise's experts to resolve the problem detected and take into account their proposals need a lot of time and effort, and sometimes even cause new problems because there are experts who are not specialized in this type of problem or who do not have enough experience to find a relevant solution. Indeed, a very important aspect to consider is the

* Corresponding author.
E-mail address: nawal.sadhouari@univ-usto.dz

problem of estimating the effort made for the accomplishment of tasks or projects (to better plan and direct this effort in the medium and long term according to the policy of the company). Thus, the absence of mechanisms to monitor the evolution of skills and performance in real time to make choices in new projects could lead to failure in collaborative work with experts in the field.

So the idea of this work is to take advantage of the benefits of recommendation systems to select the best qualified and competent experts and best placed to participate in the negotiation sessions. Our contribution is materialized by:

- Design and development of a tool for recommending the most competent business experts to resolve inconsistencies in business rules during negotiation sessions.
- Application of the PROMETHEE II multicriteria method to rank the recommended experts from best to worst.
- Conflict resolution by proposing a dynamic negotiation protocol based on the extended version of the CNP, where the proposal of the first expert classified by PROMETHEE II is evaluated.

The article is organized as follows: Section II presents some related works. Section III deals with the problems encountered within the SONATRACH enterprise and highlight the proposed contribution. In Section IV, our proposed approach is explained. This section is followed by a discussion of the obtained results. Finally, Section VI provides the conclusion of this paper, including potential direction for future research.

## II. Related Works

Recently, data on the web has increasingly become large, and humans can't treat them with traditional tools. Hence the need to use a recommendation system in order to filter such enormous size of information and extract only the useful part has risen. Recommendation systems are applied in several areas, such as movies, music, books, and so on.

### A. Recommendation Systems in Different Domains

The recent rapid growth of the Internet content has led to building recommendation systems that guide users to their needs through an information retrieving process [2]. Currently, there are three main filtering approaches: content-based, collaborative, and hybrid. Content-based filtering compares new items against each user's profile, and recommends the closest ones. Collaborative filtering compares users against each other on the basis of their past judgments to create communities, and each user receives the items deemed relevant by their community. Hybrid filtering combines content-based filtering and collaborative filtering to make the most of each other's advantages [3]. In what follows, the most recent work using recommendation systems in different fields are presented.

Authors in [4] proposed to use the Linked Open Data which is a publicly available set of interlinked data and documents, in order to find enough information about new items.

A survey is proposed in [5] that presents the phases of recommendation process and explores the different recommendation filtering techniques and the evaluation metrics for recommendation algorithms.

In [6], authors proposed a recommendation system based on two collaborative filtering algorithms in order to enhance the prediction accuracy in the big data context. The first algorithm uses the k-means clustering technique while the second one uses the k-means clustering technique coupled with Principal Component Analysis.

The paper in [7] described an approach that combines linked data cloud and the information filtering process using a semantic space vector model, and FOAF vocabulary, to define a new distance measure between users.

A new approach has been proposed in [8] to resolve the new user problem in collaborative filtering recommender systems. Authors analyze three solutions, to address the new user cold-start problem, based on the exploitation of user personality information, namely: personality-based collaborative filtering, personality-based active learning and personality-based cross-domain recommendation.

Another approach that addressed the cold-start recommendations and content-based recommendation has been proposed in [9]. Authors presented an optimization model for extracting the relationship hidden in content features by considering user preferences. The method was tested on three public datasets that are: hetrec-movielens-2k-v2; book-crossing and Netflix.

The paper in [10] proposed a system which recommended movies by using data clustering and nature-inspired algorithm. The K-means algorithm is used for clustering with nature-inspired algorithm in order to achieve a global optimum solution.

Other authors proposed a recommendation process for auto industry based on collaborative filtering and association rules. They used association rules in order to classify and find potential customers, then they applied the collaborative filtering methods to realize recommendations [11].

Another work [12] combined an implicit social graph, association rules and pairwise association rules in order to implement a recommender algorithm for food.

The paper in [13] explored different ways of combining predictions from the two types of collaborative filtering: User-based and item-based collaborative filtering. Authors proposed to fuse predictions through multiple linear regression and support vector regression models. The proposed approach aimed to minimize the overall prediction error.

In [14], authors proposed a new soft computing method based on machine learning techniques in order to find the best matching eco-friendly hotels based on several quality factors in TripAdvisor. A dimensionality reduction and prediction machine learning techniques is used to improve the scalability of prediction from the large number of users' ratings. To find the important features of eco-friendly hotels for users, the CART technique was used as a feature selection technique and ANFIS as a supervised machine learning technique.

LOOKER, a mobile recommender system for tourism domain was proposed in [15]. A content-based filtering strategy was implemented to make personalized suggestions based on the user's tourism-related user-generated content diffused on social media.

In [16], a recommendation system was proposed for the recommendation of movies based on the genres. A content-based filtering approach using genre correlation was presented based on the type of genres that the user might prefer to watch.

In the work presented in [17], a recommendation system for financial planning was described, using a hybrid approach that combined the user–user and item–item similarity with demographic filtering.

A personalized Context-Aware Hybrid Travel Recommender System was presented in [18], using user's contextual information. The proposed system was evaluated on the datasets of Yelp and TripAdvisor.

### B. Expert Recommendation Systems

Recommender systems have been also used to recommend experts. An expert recommendation system is an emerging area that attempts to detect the most knowledgeable people in some specific topics. This detection is based on both the extracted information from peoples'

activities and the content of the documents concerned with them. Moreover, an expert recommendation system takes a user topic or query and then provides a list of people sorted by the degree of their relevant expertise with the given topic or query. These systems can be modeled by information retrieval approaches, along with search engines or a combination of natural language processing systems [2].

The work in [19] presented an architecture based on the expertise of users and clustering. The proposed architecture is composed of: ER client, Web Browser, profiling supervisor, Profile DB, Identification Supervisor, Selection Supervisor, Prefs DB, Interaction Management and HTTP Server.

In the software engineering fields, authors in [20] described a novel expert recommendation system that is based on machine-learning algorithms and domain ontology, in order to identify individuals who could be involved in tackling new design concerns.

The objective of the approach proposed in [21] is to develop an expert recommender system based on social bookmarking systems and folksonomies, in order to find possible colleagues for establishing communities of practice, where people share the same interests and support each other in their working or scientific field. This expert recommender system used the Dice similarity and clustering to recommend similar users based on same bookmarks and tags within social bookmarking systems.

The paper in [2] presented a state of art on expert recommendation systems and explained in details the basic elements and procedures of these systems. It gave some real examples of their applications.

In order to recommend experts who have the appropriate knowledge with regards to the user information needs and detecting experts' communities in a social network, the authors of [22] proposed a hybrid recommender system that integrates the content-based characteristics into a social network-based collaborative filtering system. The proposed approach used Bag of Words model, semantic social network and k-means clustering algorithm.

In [23], an expert recommender system is proposed for the National Industry Association. The whole architecture is composed of: Data collector module, Matching modules, Storage modules, Database connection and Web service.

Furthermore, the proposed method presented in [24] aims at recommending colleague in Expert Cloud based on the friend-of-friends (FOF) concept and the All Possible Colleagues at First (APCF) method. In order to find all colleagues who are related to the target user, several features are considered like reputation, expertise, trust, cost, agility and field of study.

In order to develop a personalized expert-based recommender system, authors in [25] used C-SVM algorithm and compared the obtained results with k-Nearest Neighbor algorithm.

### C. Comparison of the Related Works

In Table I, we present a comparison of some related works that treat the recommendation.

TABLE I. Comparison of Some Related Works

| Work | Recommended Items and Services | Type of the system | | Limits |
|---|---|---|---|---|
| | | Content based filtering | Collaborative filtering | |
| [4] | Movies | | X | • Test of other similarity measures |
| [6] | Movies | | X | • Testing of K-means algorithm only |
| [7] | Movies | X | | • Test of other similarity measures |
| [8] | Movies, music and books | | X | • The not selected items are automatically labeled as dislikes |
| [9] | Movies and books | X | | • User profile features are not taken into account |
| [10] | Movies | X | | • Testing of K-means algorithm only |
| [11] | Auto industry | | X | • The number of closest neighbor sets (K = 3) |
| [12] | Foods | | | • No analyze of dietary specificities of regions |
| [13] | Movies | | X | • Small Dataset |
| [14] | Hotels | | X | • Test of other clustering techniques |
| [15] | Food, shopping, health and attractions | X | | • No privacy and confidentiality of content user<br>• No analysis of the textual content |
| [16] | Movies | X | | • Testing the Euclidian distance only<br>• Security issues |
| [17] | Financial planning | X | X | • More information about the user should be taken into account |
| [18] | E-Tourism | X | X | • Testing the Pearson Correlation only |
| [19] | Experts | | X | • No performance evaluation of the proposed system is given |
| [20] | Software engineering experts | | X | • No assignment of roles to experts<br>• Failure to take into account the availability of experts when creating the list of experts |
| [21] | Colleagues | | X | • Test of other similarity measures<br>• Use of a single threshold value (0.1) |
| [22] | Experts | X | X | • Extraction of the semantic knowledge from Wikipedia articles |
| [23] | Experts | X | X | • No data confidentiality |
| [24] | Colleagues | X | X | • The stages number of colleagues for the target user (=5) |
| [25] | Experts | X | X | • The personal expertise feature is not taken into account.<br>• Testing of KNN and C-SVM algorithms only |

## III. Addressed Problems and Contribution

### A. ExpRules Description

In the previous work, authors have proposed an agent-based collaborative system dedicated to capitalizing knowledge, experiences, skills and expertise of experts in the form of business rules, using a collaborative editor (ExpRules) [1].

The main objective behind this approach was to improve business rule consistency management and maintaining rules security without degrading system response time and performance. Domain ontology has been constructed as a formal model to represent a structured conceptual vocabulary that is used on one hand to express business rules, and on the other one to check the inconsistencies that can be detected on business rules [1].

Using the business language, the experts have the possibility to express their rules in an autonomous manner. Herein, a simple rule comprises two parts namely: a condition part and an action part (See the example below). The whole process needs to pass through several steps, from the introduction of the rule until the final storage in the rules base [26], [27], [28].

> **Consistent rule**: If priority is 2 then start the work at the next scheduled stop
>
> **Inconsistent rule**: If priority is 2 then start the work the following day of the request.

The obtained results during the experiments were very encouraging. This permits to convince the experts and senior responsible in SONATRACH to use and generalize our system.

### B. Problem Statement

In the event that an inconsistent rule is detected, the system sends a notification to the concerned expert in order to correct his inconsistent rule. If the system does not receive a response from the concerned expert, then it sends the rule and inconsistency details to the other experts and then initiates a negotiation session to fix the inconsistent rule.

During the negotiation, the system launches collaboration between all the experts of the enterprise in order to find a solution in agreement. Two scenarios are possible, either the experts in collaboration agree on the decision to be made regarding the correction of the detected error, or they find themselves in a conflicting situation, Here, the system adopts the strategies of negotiation to solve the problem, based on the CNP (Contract Net Protocol) with the extended version.

After testing ExpRules, we found that during the negotiation phase, there are experts who propose relevant and correct solutions while other experts do not even participate in the negotiation session or else they propose incorrect solutions, which increases the number of negotiation rounds and weighs down the system. This problem arises when the expert who is not qualified or who does not have enough experience to solve a problem intervenes in the process of managing inconsistencies.

On the other hand, the fact of inviting all the experts of the company to the negotiation session takes a lot of time, since each expert is specialized in a specific area, then inviting an expert who has no connection with the current issue or he/she is a little far will cause a more inconsistent problem, more effort and more time to find a solution.

Another problem raised during the test of ExpRules, is that in the negotiation stage, the first received response is evaluated while sometimes it is not the best solution, which leads to another negotiation round at least.

So the idea behind this article is to exploit the recommendation to guide the negotiation, the interest with this approach lies in the saving of time in the negotiating rounds. The participants in the negotiation will be the experts who have been recommended according to their skills and their interventions (successfully) in the previous sessions.

### C. Our Contribution

Given the large number of rules used in companies, our goal was to create a system that can detect and manage business rule inconsistencies in a very short time, following a rigorous control strategy involving the opinion of the most experienced experts in most situations.

The main objective of this work is to find a way to measure skills and assess the performance of experts in real time. These domain experts intervene during collaborative work for the management of business rules. Our main goal is to provide a tool to facilitate and better manage the inconsistencies that can be detected in the business rules, thus being able to assess the performance of experts to measure their effectiveness which can help in the evaluation of collaborative work. The interest of this tool is to establish a list of favorites among the business experts in order to make better and more thoughtful choices of people chosen in new projects. This procedure falls within the scope of the recommendation because our tool provides lists of business experts who are able to solve problems and provide new solutions based on their skills, all to reduce the response time and to have the right people in the right place and at the right time.

In this paper, we will present a new approach to manage inconsistencies in business rules through the use of dynamic negotiation, recommendation, unsupervised classification and the PROMETHEE II method, to find the most competent and similar experts, and group them in the same cluster, then we apply the PROMETHEE II method intra cluster. To do this, we based on the skills of business experts and their efficiency in interventions during collaborative work to manage business rules.

We summarize our main contribution in the following points:

- Collecting experts' preferences and skills explicitly and implicitly, in order to evaluate their performance and measure their effectiveness,
- Applying an unsupervised classification algorithm in order to classify and recommend experts,
- Applying PROMETHEE II to deal with the problem of evaluating the first response,
- Applying dynamic negotiation to resolve inconsistencies in business rules.

## IV. Proposed Approach

This paper presents a new approach to measure the skills of business experts in companies to assess their performance in real time, these experts intervene during the collaborative work for the business rules management. The paper presents an approach that helps to recommend business experts with high qualification and expertise in consistency management rules. These experts are intervening during collaboration and negotiation sessions to detect and correct business rules in maintenance field. The main advantage of this suggested idea is to take benefit from the integration of the recommender tool, unsupervised clustering and multi-criteria decision support methods in the knowledge based system ExpRules.

Fig. 1 presents the proposed architecture which is composed of:

- Collaborative Knowledge-based System: that allows the experts to introduce, manage and update their business rules using a

domain ontology. The domain ontology is used to detect problems of inconsistency detected in the introduced rules and store the entities used in the edition of business rules [1].

- Recommender tool: that provides a list of competent experts who can solve the detected problems. In this paper, we will focus on this phase.

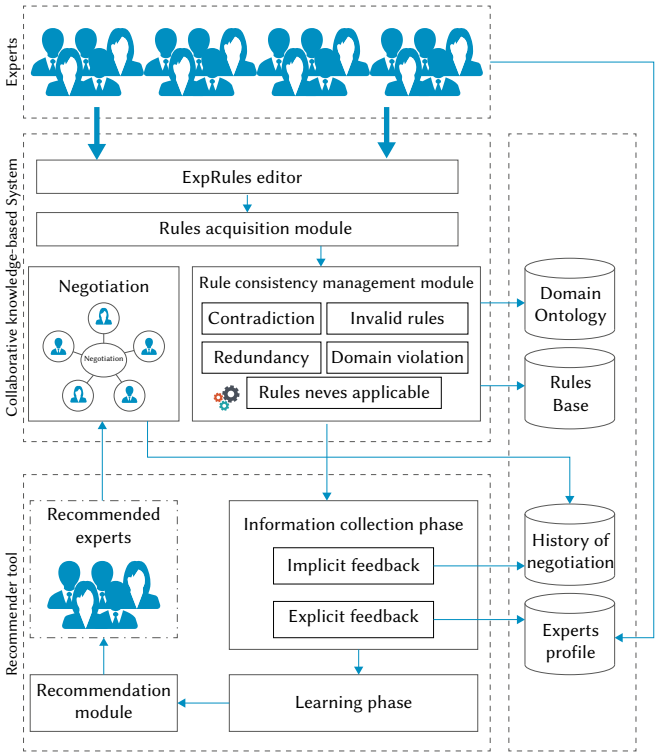



Fig. 1. Architecture of the proposed approach.

## A. Process for Recommending Experts

In order to better understand how our recommendation process for business experts works, we propose to follow four sequential steps, which are:

### 1. Rule Introduction

This is the first step of the process that represents the introduction of the rule by the business expert, this rule is usually introduced in the form "IF Conditions THEN Actions". The premises are described in the IF section of the rule and represent facts or situations and the conclusions are described in the THEN section [26].

### 2. Consistency Verification

The problem of the consistency management of the business rules defined by the experts is a very difficult problem. A business rule management system must ensure that all business rules include only those rules that are consistent and do not conflict with one another [29]. Our system addresses the following inconsistencies: contradiction, never-applicable rules, invalid rules, domain violation, and redundancy [26].

### 3. Experts' Recommendation

After the inconsistencies detection of the introduced rule, a list of recommended experts is obtained as follows:

#### a) Information Collection Phase

This phase collects relevant information of experts to generate an expert profile for the recommendation tasks including user's attribute, behaviors or content of the consistency problem the expert accesses. In our system, we use explicit and implicit feedback as follows:

#### Explicit Feedback

In its first registration, each expert fills a questionnaire to select the types of inconsistencies that can resolve (see Table II).

TABLE II. Expert Preferences Retrieved from the Questionnaire

| | Invalid rule | Never-applicable rule | Contradiction | Domain violation |
|---|---|---|---|---|
| **Expert 1** | Yes | Yes | Yes | Yes |
| **Expert 2** | No | No | No | Yes |
| **Expert 3** | Yes | Yes | Yes | No |
| **Expert 4** | Yes | Yes | Yes | No |
| **…** | | | | |
| **Expert n** | Yes | Yes | No | Yes |

Next, we calculate the total number of yes and no as shown in Table III.

TABLE III. Calculated Expert Preferences

| | Yes | No |
|---|---|---|
| **Expert 1** | 4 | 0 |
| **Expert 2** | 1 | 3 |
| **Expert 3** | 3 | 1 |
| **Expert 4** | 3 | 1 |
| **…** | | |
| **Expert n** | 3 | 1 |

After retrieving Table II and III, we based on the history of the rules already introduced (coherent or not) to measure the skills of the business experts as follows: an expert gets +1 when he introduces a consistent rule from the first time and -1 otherwise. Following this principle, we get the values of the Table IV.

TABLE IV. Expert Skills

| | Number of consistent rules | Number of inconsistent rules | Total |
|---|---|---|---|
| **Expert 1** | 6 | -2 | 4 |
| **Expert 2** | 1 | -7 | -6 |
| **Expert 3** | 4 | -3 | 1 |
| **Expert 4** | 1 | -4 | -3 |
| **…** | | | |
| **Expert n** | 0 | -1 | -1 |

Then we calculate the total for each expert, taking into account his preferences and his skills. The finality of this step is a list of experts participating in the implicit feedback step. In this step, if an expert takes a value <=0 or the total number of "yes" = 0, he is excluded from this list.

#### Implicit Feedback

The system automatically infers the expert's preferences by monitoring the different actions of expert such as the history of negotiation.

The second step is based on the negotiation history. In this step, we retrieve all previously resolved rules with their detected inconsistencies and the experts who solved the problem (see Table V).

TABLE V. The History of Previous Rules

| | Invalid rule | Never-applicable rule | Contradiction | Domain violation |
|---|---|---|---|---|
| **Rule 1** | X | X | X | |
| **Rule 2** | X | X | | X |
| **Rule 3** | | | | |
| **Rule 4** | | | X | |
| **…** | | | | |
| **Rule n** | X | X | | X |

### *b) Learning Phase*

In this phase, we apply a learning algorithm to filter and exploit the expert's features from the feedback gathered in information collection phase.

#### *Expert Weight Recovery*

After retrieving the negotiation history, we take the inconsistent rules and we recover the experts who have proposed a solution to solve these problems. Once the list of experts is established, we fill in Table VI which represents the number of the inconsistency i solved by the expert j.

TABLE VI. Expert Weights

| | Invalid rule | Never-applicable rule | Contradiction | Domain violation |
|---|---|---|---|---|
| **Expert 1** | 6 | 7 | 1 | 1 |
| **Expert 2** | 0 | 0 | 0 | 9 |
| **Expert 3** | 10 | 11 | 3 | 0 |
| **Expert 4** | 17 | 8 | 4 | 0 |
| **…** | | | | |
| **Expert n** | 14 | 5 | 0 | 8 |

#### *Expert Classification*

After that, we apply an unsupervised classification algorithm to group the experts in clusters (See section 4.B). Based on expert profiles and negotiation history, the system searches for experts who are the most similar. To do this, we used two algorithms, K-means and Fuzzy C-means to compare them and find the most suitable algorithm for our case.

### *c) Recommendation Phase*

In this phase, a list of recommended experts is proposed. Once the clusters have been generated, we select the cluster which contains the most suitable experts to solve the problem encountered. To do this, for each cluster, we calculate the number of experts who resolved the same inconsistencies as the rule in question. Next, we calculate the total of the inconsistencies resolved in each cluster, and the rule is assigned to the cluster that has the highest number of the inconsistencies resolved.

## *4. Negotiation*

After having establishing the list of the recommended experts, a message containing the inconsistent rule as well as the evaluation report, will be sent to each expert on the list to ask for their help, here a collaboration session starts, which is aimed at solving the problem found in the business rule (See section 4.C).

## *B. Unsupervised Classification*

The use of unsupervised classification allows us to reduce the load and the response time necessary for the detected problems resolution, through the formation of communities which allows us to launch the negotiation only between the users belonging to the same community.

To classify the experts, we use the K-means and Fuzzy C-means algorithm to compare them and choose the best algorithm and the most suitable for our case.

### *1. K-means Algorithm*

K-means is a non-hierarchical unsupervised clustering algorithm. It allows the observations of the data set to be grouped into K separate clusters. Thus similar data will be found in the same cluster. In addition, an observation can only be found in one cluster at a time (exclusive membership). The same observation cannot therefore belong to two different clusters [30].

The k-means algorithm is the best known and most used clustering algorithm, due to its simplicity of implementation [8]. We chose the K-means algorithm because it is efficient, simple to implement and scalable, given its ability to process very large databases and only the vectors of the means are to be kept in main memory, in addition to its linear complexity relative to the number of observations.

The pseudo code of the K-means is presented in algorithm 1.

---

**Algorithm 1**: K-means

**Input**:

- K the number of clusters to be formed
- The Training Set

**Output:** K clusters

**Begin**

1. Randomly choose K points (experts). These points are the centers of the clusters (named centroïd);

**REPEAT**

- Assign each expert in the data matrix to the group of which he is closest to his center;
- Recalculate the center of each cluster and modify the centroid;

**UNTIL** (CONVERGENCE)

**End**

---

To be able to group a dataset into K separate clusters, the K-Means algorithm needs a way to compare the degree of similarity between the different observations. Thus, two data which are similar, will have a reduced dissimilarity distance, while two different objects will have a greater separation distance. Equation (1) shows the Cosine similarity measure used.

$$Cosine\ similarity = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \quad (1)$$

### *2. Fuzzy C-means Algorithm*

The fuzzy C-means algorithm is a fuzzy unsupervised classification algorithm, which is based on the same principle as K-means but which uses the logic of fuzzy sets (use of probabilities). Fuzzy C-means is a method of clustering which allows one piece of data to belong to two or more clusters [31].

We opted for the fuzzy unsupervised classification because an expert can belong to several clusters with a certain degree of belonging. We chose the Fuzzy C-means algorithm because of its simplicity and popularity, and it is considered among the best performing fuzzy algorithms.

The pseudo code of Fuzzy C-means is presented in algorithm 2.

**Algorithm 2**: Fuzzy C-means

**Input**:

- K the number of clusters to be formed
- The Training Set
- M: degree of fuzziness
- $\varepsilon$ epsilon
- U: matrix to be initialized with random values in the interval [0.1]

**Output**: K clusters

**Begin**

1. Initialize the centers;
2. Set the parameter m (fuzzy coefficient);
3. Calculation of the initial fuzzy partition U (the membership matrix);

**REPEAT**

- Calculation of new centers ;
- Calculation of the new fuzzy partition;

**UNTIL** (CONVERGENCE)

### C. Negotiation

In order to detect and manage inconsistencies in business rules in a very short time, we have grouped similar experts in clusters to launch dynamic negotiation only between experts in the same cluster. The idea behind the use of clustering in negotiation is to save time and above all to avoid the participation of experts which cannot provide a solution to the problems detected and therefore weigh down the system.

After selecting the cluster containing the most suitable and competent experts who can solve the problems of the introduced rule, we apply a multi-criteria analysis method by partial aggregation, namely "PROMETHEE II" inside the chosen cluster. PROMETHEE II is a multi-criteria method which makes it possible to resolve the ranking problem in order to classify all the experts in the cluster from "best" to "worst".

The criteria weights used in the PROMETHEE II method are presented in Table VII.

TABLE VII. Criteria Weights of PROMETHEE II

| Criteria | | Weights |
|---|---|---|
| Introduction of consistent rules | | 0.3 |
| Introduction of inconsistent rules | Invalid, domain violation or not applicable rule | 0.1 |
| | Contradiction | 0.2 |
| Intervention in problem solving with coherent solutions | | 0.3 |
| Intervention in problem solving with inconsistent solutions | | 0.1 |

Once the PROMETHEE II method has been applied, we send the introduced rule and its corresponding evaluation report to the selected experts, and then we wait for their responses. After the deadline is over and the various responses are collected, several scenarios can occur. In the following we will present the most important scenarios:

- If all the experts decide to delete the rule, then the rule will be deleted.
- If the experts send modifications of the introduced rule, then the system will evaluate the consistency of the expert's response, which is ranked first (by the PROMETHEE II method) in its cluster. If his rule is inconsistent then the system will evaluate the response of the expert who is ranked second, otherwise the new rule will be sent again to the other experts in the same cluster. These scenarios will be repeated until convergence and the joint agreement of all the experts in the cluster.

## V. Implementation and Discussion

We developed our application and launched the simulations on an Intel (R) Core (TM) i7-3600M CPU with a speed of 3.20 GHZ, with a memory capacity of 8.00 GB of RAM under Windows 10.

### A. A Simple Scenario Illustration

When starting the tool, a main window will appear; this latter gives the possibility to authenticate according to the type of profile. In this work, we have two types of profiles: Expert and Administrator.

In what follows, we consider a simple scenario to illustrate our approach. An expert wants to introduce the following business rule: "If priority is 2 then start the work at the next scheduled stop". The Fig. 2 shows the interface which allows an expert to introduce the rule.

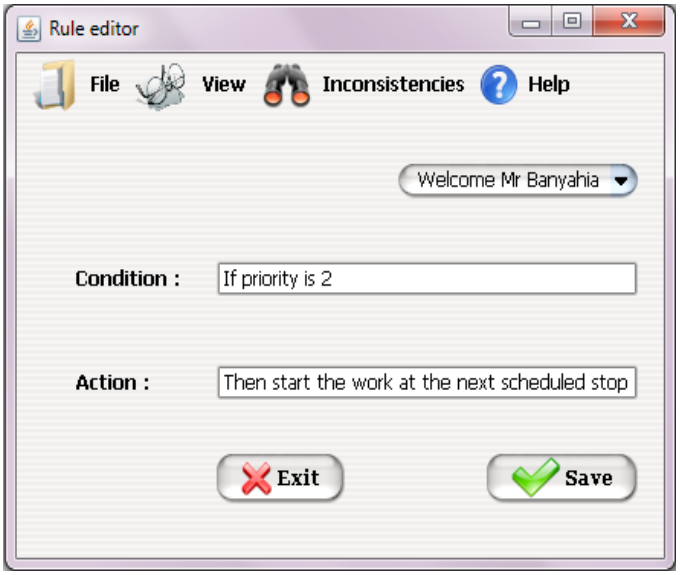



Fig. 2. Introduction of a business rule.

While checking the consistency, the system detects that the entered rule is inconsistent, so it will be stored in a temporary rule base and a notification is sent to the expert with a detailed description of the detected problem.

If the expert does not respond after two days, then a recommendation list of the most competent experts is proposed taking into account their preferences as well as the trading history in order to launch the negotiation and solve the problem (Fig. 3).

The system sends the incoherent rule to the recommended experts with a report which describes the encountered problem in this rule. Each expert sends a response. The system collects the answers, and makes a decision on the basis of its analysis of the responses.

### B. Experimentations

In what follows, we will present the results of the experiments made to validate the proposed approach.

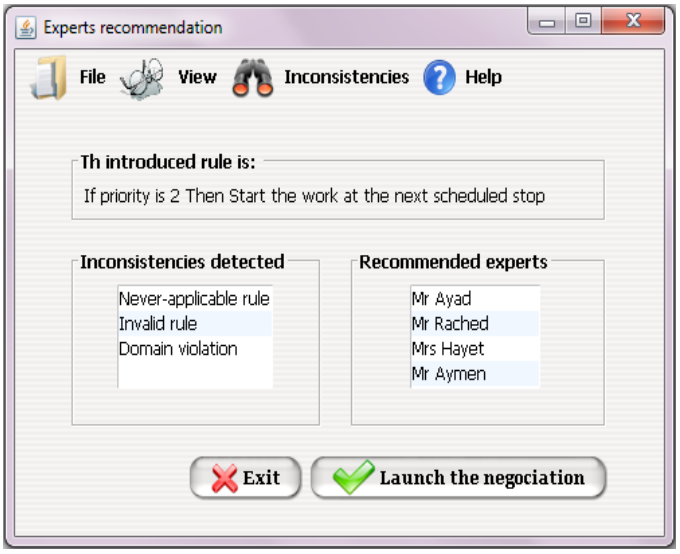



Fig. 3. The experts' recommendation.

### 1. Experiment 1: Choice of the Number of Clusters

Choosing a number of clusters K is not necessarily intuitive. Especially when the dataset is large and it is difficult to visualize the data to determine the ideal number of clusters. A large number of K can lead to overly fragmented data partitioning. This will prevent the discovery of interesting patterns in the data. On the other hand, a too small number of clusters, will lead to having, potentially, too generalist clusters containing a lot of data. In this case, there will be no interesting patterns to discover. The difficulty therefore lies in choosing a number of clusters K which can allow experts to be grouped into significant groups [30]. In the literature, several different methods of estimating the adequate number of clusters are proposed.

The Thumb Rule defined by equation (2) is proportional to the number of points n [32].

$$K \approx \sqrt{n/2} \tag{2}$$

By applying equation (2), we find that K = 4.47, so the number of clusters suitable for our dataset is either 4 or 5.

The most widely used method for choosing the number of clusters is the elbow method which consists of running the algorithm with different values of K and calculating the variance of the different clusters. The variance is the sum of the distances between each centroid of a cluster and the different observations included in the same cluster [30].

So, we draw a graph with the experimental number of clusters on the abscissa, and the variance on the ordinate, and the best k estimated by this method is at the location of the curve where an elbow is formed [32].

The variance of the clusters is calculated as follows [30]:

$$V = \sum_{j} \sum_{x_i \to c_j} D(c_j, x_i)^2 \tag{3}$$

Where:

$c_j$: The center of the cluster

$x_i$: The $i^{th}$ observation in the cluster having centroid $c_j$.

$D(c_j,x_i)$: The Euclidean distance between the center of the cluster and the point $x_i$.

By testing several values of k, we obtain the results shown in Table VIII.

The results of Table VIII in graphical form are shown in Fig. 4.

TABLE VIII. Choice of Cluster Number

| Number of cluster | K-means Variance | Fuzzy C-means Variance |
|---|---|---|
| **K = 2** | 2.223926 | 1,82 |
| **K = 3** | 1.761421 | 2,13 |
| **K = 4** | 1.6334496 | 2,74 |
| **K = 5** | 1.4772046 | 3,05 |
| **K = 6** | 1.4442943 | 3,28 |
| **K = 7** | 1.4219319 | 3,47 |
| **K = 8** | 1.4122691 | 3,68 |

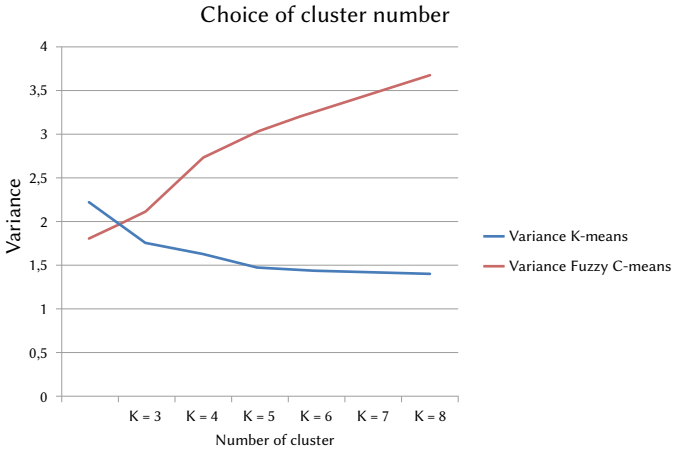



Fig. 4. Choice of the number of clusters.

For the K-means, we notice on the graph, the shape of an arm where the highest point represents the shoulder and the point where K is 8 represents the hand (the opposite for the Fuzzy C-means).

The optimal number of clusters is the point representing the knee. Here the bend can be represented by K = 5 for the K-means and K = 4 for the Fuzzy C-means.

### 2. Experiment 2: Comparison Between K-means and Fuzzy C-means

This paper proposes an approach based on clustering to manage the inconsistencies detected in the introduced business rules. The grouping of experts in clusters allows us to save time in resolving inconsistencies since only the most competent and experienced experts are invited to the negotiation session. To do this, we used the K-means algorithm and the Fuzzy C-means algorithm. Table VIII presents a comparison between the K-means and the Fuzzy C-means, in terms of recall (see equation (4)), precision (see equation (5)) and F-measure (see equation (6)).

$$Recall = \frac{Number\ of\ competent\ experts\ recommended}{Total\ number\ of\ competent\ experts} \tag{4}$$

$$Precision = \frac{Number\ of\ competent\ experts\ recommended}{Total\ number\ of\ recommended\ experts} \tag{5}$$

$$F\text{-}measure = 2 * \frac{(Precision * Racall)}{(Precision + Racall)} \tag{6}$$

We launched 17 experimentations and for each one we calculated the recall, precision and F-measure. At the end, we calculated the average of the recall, precision and F-measure, the obtained results are presented in Table IX.

TABLE IX. RECALL, PRECISION AND F-MEASURE

| | K-means | Fuzzy C-means |
|---|---|---|
| **Recall** | 0.86 | 0.64 |
| **Precision** | 0.92 | 0.47 |
| **F-measure** | 0.89 | 0.54 |

The graphical representation of the obtained results are shown in Fig. 5.

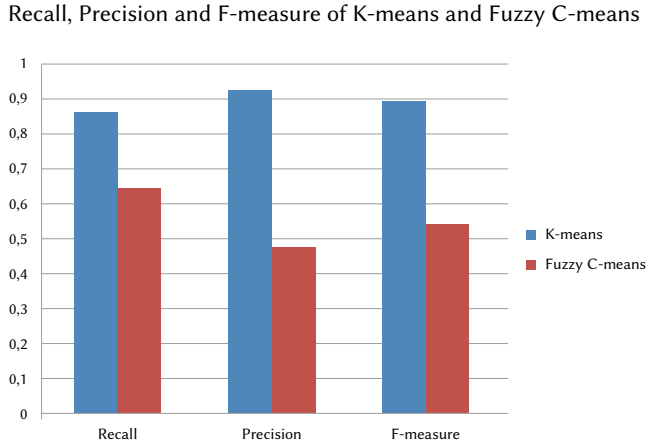



Fig. 5. Recall, Precision and F-measure.

The obtained results reveal that our system offers good results in terms of recall, precision and F-measure with the K-means algorithm.

Table X presents a comparison between the K-means and the Fuzzy C-means, in terms of response time and space memory.

TABLE X. RESPONSE TIME AND MEMORY COMPARISON OF K-MEANS AND FUZZY C-MEANS

| | K-means | Fuzzy C-means |
|---|---|---|
| **Average response time (ms)** | 50 | 54 |
| **Average space memory (bytes)** | 5445520 | 5645160 |

We note that the response time and the memory space required to form the clusters is almost the same.

By combining all the results, we can say that K-means is the most suitable algorithm for our case in terms of recall, precision, F-measure, response time and space memory.

We will use the K-means algorithm to launch the following experiments.

In order to measure the quality of the obtained clusters, we calculate the silhouette coefficient that combines ideas of both cohesion and separation. The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters. The obtained results show that most experts were very well classified.

### *3. Experiment 3: Test of Our Recommendation Tool*

In order to analyze and evaluate the behavior of our proposed approach, we measured the average response time and the average occupied memory space for 30 inconsistent rules (see Table XI).

TABLE XI. TEST OF OUR RECOMMENDATION TOOL

| | Without recommendation [1] | With recommendation |
|---|---|---|
| **Average response time (ms)** | 2446.87 | 154.28 |
| **Average space memory (bytes)** | 39954632 | 9843912 |

The aim of this experiment is to show the added value that we obtained by using the grouping with the unsupervised classification algorithm. We compared the performances of the ExpRules system which was presented in [1] and this recent new approach.

From the obtained results, we can say that the approach proposed in this paper offers significant improvements in terms of response time (15 times less than ExpRules) and memory space (4 times less than ExpRules) thanks to the use of the recommendation and unsupervised classification, that allowed us to solicit only competent experts who can provide a relevant solution to the company. The savings in response time also means savings in the effort of experts in resolving problems, which increases the company's ability to react quickly to changes.

### *4. Experiment 4: Comparing the Number of Negotiating Rounds*

We compared our proposed approach with the approach presented in [29], in terms of the number of experts invited to the negotiation session, the number of experts who participated in the negotiation session, the number of negotiation rounds and the number of inconsistent rules proposed. The results of the comparison are shown in Table XII.

From the results shown in Table XII, we can say that our approach is better compared to the approach presented in [29]. We found that the proposed approach took only one round of negotiation with no inconsistencies detected, it means that the problem was addressed from the first proposal. Also 12 experts on 12 invited experts have participated in the negotiation session. In contrast, the no-recommendation approach took 5 rounds of negotiation to come to a common agreement with the 11 inconsistent proposed rules. Also 32 experts participated in the negotiation among 40 invited experts.

TABLE XII. COMPARING THE NUMBER OF NEGOTIATING ROUNDS

| | Without recommendation [29] | With recommendation |
|---|---|---|
| **Number of experts invited to the negotiation session** | 40 | 12 |
| **Number of experts who participated in the negotiation session** | 32 | 12 |
| **Number of negotiation rounds** | 5 | 1 |
| **Number of inconsistent rules proposed** | 11 | 0 |

## VI. CONCLUSION

In companies, many decisions are made every day and some decisions are made much more difficult when faced with the large amount of data or the structural complexity of the decision to be made. Recommendations are a rapidly growing area of research to help us in this decision-making process.

The major contribution of this article is materialized by the development of a recommendation tool that can be used by managers to find the skills of the experts for managing business rules, this tool makes it possible to compare the profile of experts to certain reference features. Indeed, skills management is implemented to measure the quality of the expertise offered by each expert who is involved in the collaborative process in order to respond quickly to market changes, thus improving the overall efficiency of the company.

Thus, to achieve our goal, we have proposed an approach that is composed of four steps that are: the introduction of the rule, the consistency check of the introduced rule, the recommendation of the experts and finally the negotiation step.

We started our work by integrating the inconsistency detection algorithm and used a domain ontology as a formal model to represent a structured conceptual vocabulary that is used on one hand to express the business rules, and on the other one to check for inconsistencies that can be detected on business rules, the inconsistencies that can be detected by the algorithm used are: invalid rules, never-applicable rules, conflicting rules, and redundant rules. At the end, we proposed and implemented a tool to recommend a list of favorites for business experts to make choices in new projects.

To save time, we have classified the most similar competent experts in the same cluster using K-means algorithm, and then we have applied the PROMETHEE II method in order to launch the negotiation inside the cluster and evaluate the solution provided by the most competent expert. This allowed us to improve the performance of the proposed system and encourage experts to participate in the process of business rules inconsistencies managing.

The new proposed approach brings a lot of improvement in terms of recall, precision, response time, memory space compared to the previous approach. In fact, the big improvement is in negotiation, which aims to deal with inconsistencies in business rules. With the use of recommendation, unsupervised classification and the PROMETHEE II method, we were able to reduce the expert workload because each expert is called upon to solve problems only in their area. We have also been able to significantly reduce the trading rounds and consequently the number of incoherent proposed business rules.

For possible extensions and improvements of our present work, we propose:

- Address other inconsistency issues such as equivalency,
- Consider other skills of the business experts,
- Weight the inconsistencies because there are inconsistencies that are more important than others.
- Provide a mobile application for the notification of business experts.

## References


[1] N. Sad houari, "Conception et réalisation d'un système collaboratif pour les experts métier à base d'agents et des algorithmes de cryptage", PhD thesis, University of Oran1 Ahmed Ben Bella, Oran, Algeria, 2017.

[2] N. Nikzad–Khasmakhi, M.A. Balafar and M. Reza Feizi–Derakhshi, "The state-of-the-art in expert recommendation systems", Engineering Applications of Artificial Intelligence, vol. 82, pp. 126–147, 2019, doi: https://doi.org/10.1016/j.engappai.2019.03.020

[3] A.T. NGUYEN, « COCoFil2 : Un nouveau système de filtrage collaboratif basé sur le modèle des espaces de communautés », PhD thesis, University of Joseph Fourier, Grenoble I, France, 2006.

[4] H. Zitouni, S. Meshoul and A. Kadi, "Toward a New Solution of New Item Problem in Collaborative Recommender Systems", in *Proceedings of the* international conference on Embedded & Distributed Systems, Oran, Algeria, 2017, pp. 328-339.

[5] F.O. Isinkaye, Y.O. Folajimi, and B.A. Ojokoh, "Recommendation systems: Principles, methods and evaluation", Egyptian Informatics Journal, vol. 16, no. 3, pp. 261–273, 2015, doi: https://doi.org/10.1016/j.eij.2015.06.005

[6] H. Zarzour, M. Soltani, F. Maazouzi and C. Chemam, "An improved collaborative filtering recommendation algorithm for big data", in *Proceedings of the* international conference on Embedded & Distributed Systems, Oran, Algeria, 2017.

[7] H. Zitouni, S. Meshoul and K. Taouche, "Enhancing Content Based Filtering Using Web of Data", in *Proceedings of* International Conference on Computational Intelligence and Its Applications, Springer, Oran, Algeria, 2018.

[8] I. Fernández-Tobías, M. Braunhofer, M. Elahi, F. Ricci and I. Cantador, "Alleviating the new user problem in collaborative filtering by exploiting personality information, User Modeling and User-Adapted Interaction", User Modeling and User-Adapted Interaction, vol. 26, no. 2, pp. 221-255, 2016, doi: 10.1007/s11257-016-9172-z

[9] H. Zhang, Y. Sun, M. Zhao, T.W. S. Chow and Q. M. J. Wu, "Bridging User Interest to Item Content for Recommender Systems: An Optimization Model", IEEE Transactions on Cybernetics, pp. 1-13, 2019, doi: 10.1109/TCYB.2019.2900159.

[10] S. P. Singh and S. Solanki, "A Movie Recommender System Using Modified Cuckoo Search", in *Proceedings of* Emerging Research in Electronics, Computer Science and Technology, Lecture Notes in Electrical Engineering, Springer Singapore, 2019, pp. 471-482.

[11] L. Yao, Z. Xu, X. Zhou, and B. Lev, "Synergies Between Association Rules and Collaborative Filtering in Recommender System: An Application to Auto Industry", in García Márquez F., Lev B. (eds) Data Science and Digital Business. Springer, Cham, 2019, pp. 65-80.

[12] T. Osadchiy, I. Poliakov, P. Olivier, M. Rowland and E. Foster, "Recommender system based on pairwise association rules", Expert Systems with Applications, vol. 115, pp. 535-542, 2019, doi: https://doi.org/10.1016/j.eswa.2018.07.077

[13] P. Thakkar, K. Varma, V. Ukani, S. Mankad and S. Tanwar, "Combining User-Based and Item-Based Collaborative Filtering Using Machine Learning", in Satapathy S., Joshi A. (eds) Information and Communication Technology for Intelligent Systems. Smart Innovation, Systems and Technologies, Springer, Singapore, vol. 107, 2019, pp. 173-180.

[14] M. Nilashi, A. Ahani, M. D. Esfahani, E. Yadegaridehkordi, S. Samad, O.Ibrahim, N. Mohd Sharef and E. Akbari, "Preference learning for eco-friendly hotels recommendation: A multi-criteria collaborative filtering approach", Journal of Cleaner Production, vol. 215, pp. 767-783, 2019, doi: https://doi.org/10.1016/j.jclepro.2019.01.012.

[15] S. Missaoui, F. Kassem, M. Viviani, A. Agostini, R. Faiz and G. Pasi, "LOOKER: a mobile, personalized recommender system in the tourism domain based on social media user-generated content", Personal and Ubiquitous Computing, vol. 23, pp. 181–197, 2019, doi: https://doi.org/10.1007/s00779-018-01194-w.

[16] S. Reddy, S. Nalluri, S. Kunisetti, S. Ashok and B. Venkatesh, "Content-Based Movie Recommendation System Using Genre Correlation", Smart Intelligent Computing and Applications, vol. 105, pp. 391-397, 2019, doi: https://doi.org/10.1007/978-981-13-1927-3.

[17] N. Pereira and S.L. Varma, "Financial Planning Recommendation System Using Content-Based Collaborative and Demographic Filtering", Smart Innovations in Communication and Computational Sciences, vol. 669, pp. 141-151, 2019, doi: https://doi.org/10.1007/978-981-10-8968-8_12.

[18] R. Logesh and V. Subramaniyaswamy, "Exploring Hybrid Recommender Systems for Personalized Travel Applications", Cognitive Informatics and Soft Computing, vol. 768, pp. 535-544, 2019, doi: https://doi.org/10.1007/978-981-13-0617-4_52.

[19] D. W. McDonald and M.S. Ackerman, "Expertise Recommender: A Flexible Recommendation Architecture", in Proceedings of the 2000 ACM Conference on Computer-Supported Cooperative Work (CSCW '00), Philadelphia, PA, 2000, pp. 231-240.

[20] M. Bhat, K. Shumaiev, K. Koc, U. Hohensteiny, A. Biesdorfy and F. Matthes, "An expert recommendation system for design decision making Who should be involved in making a design decision?", In Proceedings of 2018 IEEE International Conference on Software Architecture, Seattle, WA, USA, 2018, pp. 158-161.

[21] T. Heck and I. Peters, "Expert Recommender Systems: Establishing Communities of Practice Based on Social Bookmarking Systems", in the Proceeding of I-KNOW 2010: 10th International Conference on Knowledge Management and Knowledge Technologies, Graz, Austria, 2010, pp. 458-464.

[22] E. Davoodi, K. Kianmehr and M. Afsharchi, "A semantic social network-based expert recommender system", Applied Intelligence, vol. 39, pp.1–13, 2013, doi: https://doi.org/10.1007/s10489-012-0389-1

[23] T. Reichling, M. Veith and V. Wulf, "Expert Recommender: Designing for a Network Organization", Computer Supported Cooperative Work, vol. 16, pp.431–465, 2007, doi: https://doi.org/10.1007/s10606-007-9055-2

[24] S. Hazratzadeh and N. J Jafari Navimipour, "Colleague Recommender System in the Expert Cloud Using Features Matrix", Kybernetes, vol. 45 no. 9, pp. 1342-1357, 2016, doi: https://doi.org/10.1108/K-08-2015-0221

[25] Y. Chung, H.W. Jung, J. Kim and J.H. Lee, "Personalized Expert-Based Recommender System: Training C-SVM for Personalized Expert Identification", in International Workshop on Machine Learning and Data Mining in Pattern Recognition, New York, USA, 2013, pp. 434-441.

[26] N. Sad Houari and N. Taghezout, "Integrating agents into a collaborative knowledge-based system for business rules consistency management", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 4, no. 2, pp. 61–72, 2016, doi: 10.9781/ijimai.2016.4210.

[27] N. Sad Houari and N. Taghezout, "A novel collaborative approach for business rules consistency management", in *Proceedings of* International conference on decision support system technology, lecture notes in business information processing, Springer, Plymouth, UK, vol. 250, 2016, pp. 152–164.

[28] N. Sad houari and N. Taghezout, "An agent based approach for security integration in Business Rules Management System", in *Proceedings of* International conference on intelligent information processing, security and advanced communication, Batna, Algeria, 2015, pp.6.

[29] N. Taghezout, N. Sad houari and A. Nador, "Negotiation model for knowledge management system using computational collective intelligence and ontology-based reasoning: case study of SONATRACH AVAL", International Journal of Simulation and Process Modelling, vol. 11, no. 5, pp. 403–427, 2016, doi: 10.1504/IJSPM.2016.079207.

[30] Y. Benzaki, "Tout ce que vous voulez savoir sur l'algorithme K-Means", Mr. Mint: 2018, Accessed : June 22, 2020. [Online]. Available: https://mrmint.fr/algorithme-k-means

[31] H. Ming-Chuan and Y. Don-Lin, "An efficient Fuzzy C-Means clustering algorithm", In Proceedings of 2001 IEEE International Conference on Data Mining, San Jose, CA, USA, 2001, pp. 225-232.

[32] V. Bouvier and P. Bellot, "Regroupement par popularité pour la RI semi-supervisée centrée sur les entités", in *Proceedings of* CORIA : conference en Recherche d'Infomations et Applications - 12th French Information Retrieval Conference, Paris, France, 2015, pp. 503-512.

N. Sad Houari

N. Sad Houari is an assistant professor at the Université des Sciences et de la Technologie d'Oran Mohamed Boudiaf, USTO-MB. She holds her doctorate thesis in IRAD at the university of Oran1 Ahmed Ben Bella in Algeria in 2017. She received her Master degree in Information System Technologies from the same university in 2013. She is a member of the EWG-DSS (Euro Working Group on Decision Support Systems) since 2016 and she is also a member of LIO (Laboratoire d'Informatique d'Oran) laboratory. Her research interests include Business rules modeling, Artificial Intelligence, Security, Multi-agents system, Knowledge management, Recommender system, Decision support system and Bioinformatics.

N. Taghezout

N. Taghezout is a professor at University of Oran1 Ahmed BenBella, Algeria. She holds her doctorate thesis in MITT at PAUL SABATIER UNIVERSITY in France in 2011. She also received another doctorate thesis in Distributed Artificial Intelligence from University of Oran1 Ahmed BenBella in 2008. She holds a Master degree in Simulation and Computer aided-design. She conducts her research at the LIO laboratory as a chief of the research group in Modeling of enterprise process by using agents and WEB technologies. Since she studied in UPS Toulouse, she became a member of the EWG-DSS (Euro Working Group on Decision Support Systems). She is currently lecturing Collaborative decision making, Enterprise management and Interface human machine design. Her seminars, publications and regular involvement in Conferences, journals and industry projects highlight her main research interests in Artificial Intelligence.

# What Causes the Dependency between Perceived Aesthetics and Perceived Usability?

Martin Schrepp[1]*, Raphael Otten[2], Kerstin Blum[1], Jörg Thomaschewski[2]

[1] SAP AG (Germany)
[2] University of Applied Sciences Emden/Leer (Germany)

Received 5 May 2020 | Accepted 12 December 2020 | Published 22 December 2020

## Abstract

Several studies reported a dependency between perceived beauty and perceived usability of a user interface. But it is still not fully clear which psychological mechanism is responsible for this dependency. We suggest a new explanation based on the concept of visual clarity. This concept describes the perception of order, alignment and visual complexity. A high visual clarity supports a fast orientation on an interface and creates an impression of simplicity. Thus, visual clarity will impact usability dimensions, like efficiency and learnability. Visual clarity is also related to classical aesthetics and the fluency effect, thus an impact on the perception of aesthetics is plausible. We present two large studies that show a strong mediator effect of visual clarity on the dependency between perceived aesthetics and perceived usability. These results support the proposed explanation. In addition, we show how visual clarity of a user interface can be evaluated by a new scale embedded in the UEQ+ framework. Construction and first evaluation results of this new scale are described.

## Keywords

Usability, Aesthetics, Visual Clarity, Layout, User Experience.

DOI: 10.9781/ijimai.2020.12.005

## I. Introduction

To be successful in today's quite competitive markets, products must be easy to use and should have an attractive and beautiful design. Research focused for a long period of time mainly on usability aspects (for example, efficiency, learnability, intuitive use, controllability or error tolerance) of products. In the last decade the focus widened to cover also user experience aspects [1], [2] (for example, aesthetical impression, stimulation or novelty). A natural question is how these usability aspects and user experience aspects relate to each other.

At first sight, beauty and usability seem to be unrelated quality aspects of a user interface, which can be designed and developed independently. But several influential studies [3], [4], [5] demonstrated that perceived aesthetics or beauty has an impact on the perceived usability of a product. This finding is often condensed in the well-known statements *What is beautiful is usable* [5] or *Attractive things work better* [6].

But the strength of the influence of perceived aesthetics on perceived usability varies between studies. Many studies found just a small influence or no effect at all [7], [8]. In addition, some authors report a reverse effect from perceived usability to perceived aesthetics (short: *What is usable is beautiful*) [9], [10], i.e. a good impression concerning the usability of a product improved the visual appeal of this product.

Thus, the effect seems to depend on different factors that vary between studies. The aesthetic impression of a user interface can be manipulated by many variables (colour of UI elements, typography, alignment, grouping, etc.).

* Corresponding author.
E-mail address: martin.schrepp@sap.com

The same is true for the usability. Quite different interaction styles can be used for the design of a user interface. In addition, the type of the investigated product may also have an impact here. For example, two recent papers [11], [12] showed that the importance of single UX aspects differ massively between product types. And of course, the importance of the UX aspects like aesthetics, learnability or efficiency has some impact on the judgement of subjects concerning this aspect and thus has an influence on whether a dependency between such ratings exists or not.

Since the effect of aesthetical impression on perceived usability or actual performance depends on so many variables, the question of how such an influence can be explained by psychological processes is quite important. A good explanation will help to understand which factors play a role and thus to predict under which circumstances we can expect a positive impact of the beauty of an interface on the perceived usability or even performance measures and under which conditions such an effect is unlikely.

Several psychological mechanisms have been proposed to explain the dependency between perceived aesthetics and perceived usability.

A popular explanation by Don Norman [6] assumes that the mood or emotional state of the user is responsible for this dependency. From psychological research we know [13] that a positive emotional state of a person improves his or her creativity and flexibility in problem solving. A negative emotional state on the other hand favours a systematic, inflexible and analytical problem-solving behaviour [14].

When interacting with a user interface, a user in a good mood should be more likely to overcome problems with creative ideas and would therefore judge them as less severe. A user in a bad mood, on the other hand, will be more focused on problematic details. Therefore, a user in a bad mood should assess the usability of a user interface

worse than a user in a good mood [6].

The basic idea behind Norman's explanation is that a beautiful design of a product causes a positive mood, while an ugly design causes a negative mood. Several papers have indeed shown that the design of a product can influence the mood of its users [15], [16]. The mood or emotional state of the user acts in this explanation as a mediator variable between perceived aesthetics and perceived usability.

A potential weakness of this explanation is that it offers no good explanation for the positive impact of perceived usability on aesthetic impression (short: *What is usable is beautiful*), which is found in two studies [9], [10] as already mentioned above.

Another often cited explanation is based on the attractiveness stereotype (the so-called HALO-effect). Several psychological studies, see for example [17], [18], have shown that people associate an attractive appearance (which is directly observable when they meet an unknown person for the first time) with other desirable, logically unrelated properties of humans, for example social competency, empathy or intelligence (which are not directly observable).

Studies in consumer research show that there is a similar effect in the judgement of products (often named *evaluative consistency*). This concept describes the tendency of people to infer missing product information from an overall evaluation of the product. For example, if a product is placed in a higher price segment often a high quality is assumed [19]. If we transfer this to user interfaces, then missing information concerning usability of a product should be inferred from the directly visible aesthetical impression of the user interface. This explanation is especially convincing if users have not interacted heavily with a product when they make their judgement, since in this state they have not much information about the quality of the interaction design and thus rely on their judgements concerning the directly visible graphical quality.

The *general impression model* [20] assumes that the overall impression of an object influences single aspects of the impression. Thus, if a user has a good overall impression of a product, he or she will also judge single aspects, for example aesthetics or usability, positively and vice versa.

A study [9] that compared both explanations could not clearly decide which one is more adequate. Both explanations were not able to explain the resulting data in this study.

In this paper, we propose a third explanation for the dependency between perceived aesthetics and perceived usability. The basic idea behind this explanation is to assume a common factor in product perception that influences both the perception of aesthetics and usability. This common factor would thus explain a dependency in both directions.

## II. Visual Clarity as Common Factor

What do we mean by the term visual clarity and why does it impact both aesthetic impression and perceived usability?

In [21] two components of aesthetic impression are distinguished. The concept of classical aesthetics describes design aspects like symmetry, clarity and order. On the other hand, expressive aesthetics focuses of creativity and originality of the design. Thus, terms like clear, clean, symmetrical, organised and ordered represent classical aesthetics, while terms like creative, original or sophisticated represent expressive aesthetics.

The VISAWI questionnaire [22], a standard questionnaire to measure visual aesthetics of web pages, contains also some items that point in the direction of classical aesthetics, for example *The layout appears well-structured.*

Many experimental papers also point in this direction. To illustrate this, we describe a few examples. In [23] it was shown that balance and symmetry of the layout improve the aesthetic impression of a design. A popular measure for layout complexity [24] uses mainly alignment of elements and variety of element sizes to calculate the complexity of a typographic layout. Results in [25] demonstrated that visual complexity and perceived order of the layout have an impact on perceived aesthetic impression and concerning preferences for websites. These results are also in line with the well-known fluency effect [26], which describes the observation that objects that are easier to process cognitively are perceived as more aesthetic. A very basic formulation of this idea dates even back to the middle of the last century. Birkhoff's *aesthetic measure* [27] uses the ratio of order and complexity to measure the aesthetic value of an object.

Thus, if we summarise these arguments, the impression of a clear, clean, structured, organised layout improves the perceived aesthetics. In the following, we call this impression visual clarity.

But items that cover this aspect of product perception can be found in other UX questionnaires as a representation of classical usability dimensions. For example, the UEQ [28], [29], [30] contains an item *organised/cluttered*, which represents the dimension *Efficiency* and an item *clear/confusing* that represents the dimension *Perspicuity* (how easy is it to understand and learn to use the product). The AttractDiff2 [31] contains an item *confusing/clear* in the scale *Pragmatic Quality* (which is merely a representation of classical usability aspects). There are many other examples of this type in other UX questionnaires. For example, the PSSUQ [32] contains an item *The organisation of the information on the systems screens was clear* as an indicator for the scale information quality. A similar statement *The website seems clearly arranged and not cluttered* is used in the NRL as part of the scale aesthetics [33].

Intuitively it is quite natural that the visual clarity of a user interface influences also usability judgements. Of course, a clear and structured user interface that contains only a small number of elements is easier to scan than a complex cluttered user interface. Thus, the time to detect the important elements for a task and thus efficiency will be influenced by visual clarity as well [34]. In addition, a high visual clarity will create the impression that the user interface is of low complexity and thus easy to learn.

It is therefore plausible to see here a simple and natural explanation for the connection between perceived usability and aesthetics. If a user interface gives a clear, well-structured impression, this should positively influence the perceived aesthetics as well as the assessment of usability. This would also explain well why there is empirical evidence for both directions (*What is beautiful is usable* and *What is usable is beautiful*).

## III. Pre-study

The goal of this study was to develop items that can be used to measure visual clarity.

### A. Participants

Participants were recruited by sending the link to the online study to a mailing list. 21 persons participated in the study (average age 29.9 years, 67% females, 33% males). Participants did not receive any benefits for their participation in the study.

### B. Material

Screenshots (size 1024 x 768 px) of the homepages of four German universities were used as stimuli. We selected pages with varying levels of complexity. Complexity was measured during the selection process by the jpeg-size of the screenshot. This is a common method to get a rough measure of complexity [35], [36], [37].

### C. Procedure

Participants could start the study over a link in the invitation mail. The first screen contained a short introduction to the study. Then the participants could navigate to a screen where they can rate the four screenshots concerning their complexity.

The students provided their subjective rating of visual complexity on a 7-point Likert scale by answering the following question:

*The homepage of the university looks*
*simple o o o o o o o complex*

The goal of this rating was to force the participants to think this concept over.

After this rating was submitted a free text question was presented. Participants were asked to list aspects of the four screenshots that are related to visual clarity or visual complexity. Finally, a second free text question about web pages in general was shown. The participants were asked to complete the sentence *"A complex web page is for me a page that ..."*.

### D. Results

Complexity ratings and jpeg-size showed that the four selected pages indeed varied sufficiently, but the order of the screenshots by perceived visual complexity does not perfectly correspond to the order by jpeg-size:

- *Page A*: perceived complexity 4.0, size 73 KB
- *Page B*: perceived complexity 4.3, size 185 KB
- *Page C*: perceived complexity 4.8, size 191 KB
- *Page D*: perceived complexity 3.4, size 128 KB

The perceived complexity represents the rating on the 7-point Likert scale described above.

The free text comments were analysed and clustered according to their semantic meaning. Concerning visual clarity two clusters emerged. One cluster contained statements concerning the number of elements on the page. The statements in the second cluster points to the perceived order and alignment of page elements, i.e. the visual organisation of the content.

### E. Conclusions

Thus, the two statements *The page has many elements* and *The information is clearly arranged on the page* were selected to represent the concept of visual clarity in the following study.

## IV. First Study

The first study tries to investigate if there is a mediator effect of visual clarity on the dependency between perceived aesthetics and perceived usability.

### A. Participants

Participants were recruited over social networks and online forums. 425 persons participated in the study. Average age of the participants was 30.77 years. 43% of the participants were males, 39% females and 18% did not provide gender information.

The dropout rate (percentage of participants that started the online-study but did not submit responses) was 40%.

### B. Material

As stimuli the start pages of 30 public German websites were used. Websites were selected from the three different categories *cities*, *web-shops* and *design agencies* to cover a broad spectrum of different cases of use and design styles. For each category a larger sample of pages (around 50) were selected. From this sample 10 pages that varied as to visual complexity (again measured by the size of the saved screen shots in jpeg-format) were selected.

For each of the 30 selected start pages a screen shot with resolution 1024 x 768 was used. Fig. 1 shows two examples of the prepared screenshots for each of the three categories.

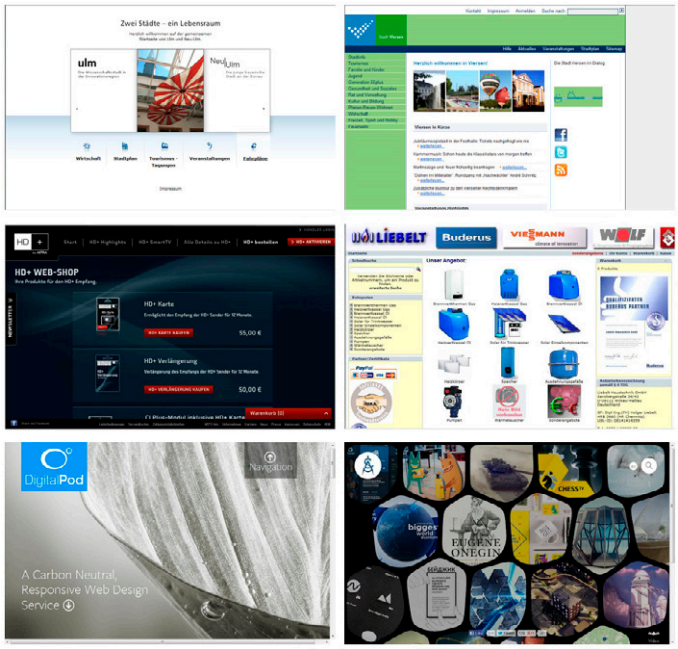



Fig. 1. Six of the used screen shots (on top two homepages of German cities, middle two start pages of web-shops and bottom two homepages of design agencies).

### C. Items

Four items were used to capture the impression of the shown pages:

- *I1: The page has many elements*
- *I2: The information is clearly arranged on the page*
- *I3: I think I would get along well with the web page*
- *I4: The design of the page is nice*

The first and second items represent the concept of visual clarity. As an indicator for visual clarity the mean value of the first two items is used. Here item one is scaled in a reverse order, since agreement to item one means a lower visual clarity. Item three is used as an indicator for the perceived usability of the pages and item four as indicator for visual aesthetics.

All items could be answered on a 7-point Likert scale with the extreme points *Do not agree at all* and *Totally agree.*

### D. Procedure

Each participant was assigned to one of the three website categories. First, a page with general instructions describing the flow of screens in the study and the tasks in each step was presented.

After the participant read this instruction, he or she could start the main part of the study over a link. A randomly selected homepage is shown as a screenshot. Below this screenshot the questions *I1* to *I4* are presented. After the participant submitted the answers the next randomly selected homepage was presented. This was repeated three times, i.e. each participant evaluated three randomly selected homepages. The restriction to three pages was meant to limit the time required to complete the study and avoid a high dropout rate.

### E. Results

The correlations between the investigated variables were highly

significant:

- *Usability, Aesthetics*: r=0.44 (t(1072)=16.24, p<0.001)
- *Clarity, Usability*: r=0.71 (t(1072)=33.39, p<0.001)
- *Clarity, Aesthetics*: r=0.51, (t(1072)=19.18, p<0.001)

The partial correlations between usability and aesthetics if the influence of clarity is controlled is 0.138 (t(1072)=4.57, p<0.001). Thus, if the impact of clarity is considered, then the dependency between the other two variables is much lower. This is a first hint that points in the direction of a mediator effect.

To clarify this in more detail we perform two mediator analyses.

First, we analyse the impact of visual clarity on the influence of aesthetics on usability. The results of the mediator analysis are depicted in Fig. 2. The values without parentheses are the regression coefficients of the simple regressions between variables, i.e. the simple regression of aesthetics on usability, aesthetics on clarity and clarity on usability.

The values in parentheses represent the regression coefficients of the combined regression of aesthetic and clarity on usability. All dependencies are significant with p<0.01).

The impact of aesthetics on usability is massively reduced if the mediator variable clarity is considered. The Sobel test [38] shows also a significant mediator effect (Sobel z = 17.05, p<0.01).

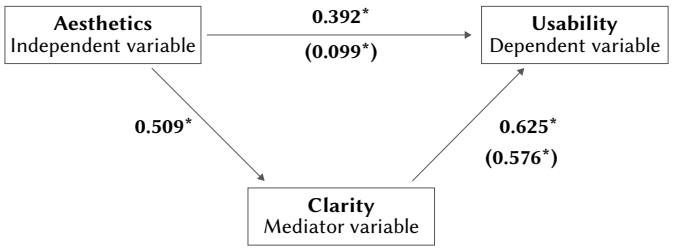



Fig. 2. Dependency between aesthetics and usability considering the impact of visual clarity.

Now we take a look at the opposite direction. The results of the mediator analysis are shown in Fig. 3. Again, there is a significant mediator effect (Sobel z = 16.78, p<0.01).

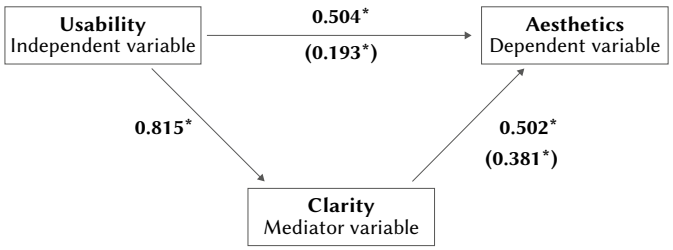



Fig. 3. Dependency between usability and aesthetics considering the impact of visual clarity.

### F. Conclusions

The study found a mediator effect of visual clarity on the dependency between perceived usability and perceived aesthetics in both directions. Thus, the results support the proposed explanation for this dependency.

However, this first study has some methodological limitations worth mentioning. First, the participants rated usability, beauty and visual clarity based on screenshots and did not interact with the pages. This will of course have an impact, especially on the judgements concerning usability.

Second, the ratings concerning usability, aesthetics and clarity were done with simple statements developed in a small pre-study that were expected to cover these concepts.

At least for usability and aesthetics there are established standard questionnaires that allow a more reliable measurement of these concepts. They were not used in this study intentionally to keep the number of items to be answered low and thus to allow the participants to rate more than one screenshot with reasonable effort. But to be able to generalise the results, a replication of the study using standard methods to operationalise these concepts would be helpful.

## V. Construction of a Clarity Scale

One of the limitations of the first study was that the concepts of usability, aesthetics and visual clarity were not measured with standard questionnaires. For usability and aesthetic impression such questionnaires are available, for the concept of visual clarity this is not the case.

In this study we describe the construction of a scale to measure visual clarity that is embedded in the UEQ+ framework [39]. The UEQ+ is a set of modular UX scales that can be combined to form a UX questionnaire. Thus, the UEQ+ allows researchers to select exactly those UX aspects as scales that are relevant for a concrete product evaluation respectively research question.

The UEQ+ is available free of charge. Scales and required material to set up a questionnaire and analyse the results can be downloaded at ueqplus.ueq-research.org.

### A. Selection of an Initial Item Set

A pool of items meant to represent the concept of visual clarity was constructed by querying several UX experts. After several discussion rounds the constructed item pool was consolidated into a candidate set of 8 items in the UEQ+ format. Thus, each item consists of a pair of terms of opposite meaning that can be rated on a 7-point Likert scale. An example is shown below:

*unorganised* o o o o o o o *organised*

The following candidate items were constructed. The German original version that is used in the study is shown in parentheses:

- *difficult to grasp / easy to grasp* (schlecht zu erfassen / gut zu erfassen)
- *poorly structured / well structured* (schlecht gegliedert / gut gegliedert)
- *unclear / clear* (unklar / klar)
- *unstructured / structured* (unstrukturiert / strukturiert)
- *disordered / ordered* (ungeordnet / geordnet)
- *unorganised / organised* (unorganisiert / organisiert)
- *ill-conceived / well-conceived* (undurchdacht / durchdacht)
- *random / planned* (zufällig / geplant)

### B. Study for Scale Construction

An online questionnaire was used to collect some response data concerning the constructed items from a larger sample.

#### 1. Participants

69 persons recruited over social media participated in the study. Average age was 29 years, 46 were males and 23 females. Participants did not receive any benefit for their participation.

#### 2. Procedure

The online questionnaire consists of four pages. The participants could navigate between these pages by two buttons labelled *Next*

and *Previous* on the bottom of the page. The last page contains just a message that thanks for participation. Data were submitted when the participant clicked on *Next* on the third page.

The first page gives some general instructions and asks for age and gender of the participant. In addition, participants are instructed only to proceed if they have already used a web shop to purchase goods online.

On page two the participants are asked to name a web shop they have already used for buying goods online. Page three contains the eight items from the set of candidate items.

### 3. Results

Most participants decided to rate Amazon.de (71%), followed by Zalando.de (14.5%) and Mediamarkt.de (7.2%). 5 other shops were just mentioned by one participant.

A factorial analysis (we used the R package psych [40]) showed that a solution with one factor fitted the data quite well (according to the scree plot and the Kaiser-Gutmann criterion). The scree plot of this solution is shown in Fig. 4.

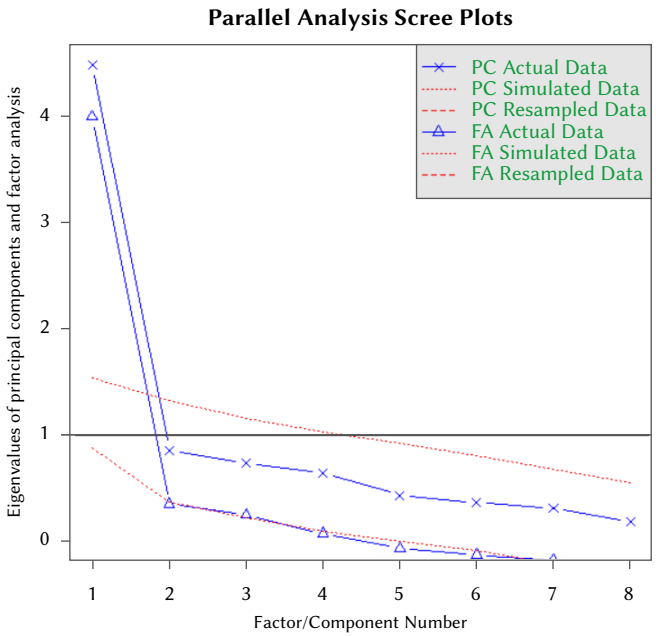



Fig. 4. Screeplot of the factorial analysis.

Thus, the four items that showed the highest loadings on this single factor were chosen to represent the scale for visual clarity (see below).

### C. Constructed UEQ+ Scale

Using the UEQ+ format the scale to measure visual clarity is:

*In my opinion the user interface of the product looks:*

| | | |
|---|---|---|
| *poorly structured* | o o o o o o o | *well structured* |
| *disordered* | o o o o o o o | *ordered* |
| *unorganised* | o o o o o o o | *organised* |
| *unstructured* | o o o o o o o | *structured* |

## VI. Second Study

The goal of this study was to replicate Study 1 with a study design that takes the limitations of this previous study into account.

A severe limitation was that the participants of study 1 just rated screenshots of web pages and did not interact with the page. Therefore, we decided to use a running web portal as stimulus and force the participants to use the main functions by giving them a task which must be solved before a rating is possible.

The quality of the rating itself is improved by using common standard questionnaires.

Usability is rated with the System Usability Scale SUS [41].

Aesthetic impression is rated with the short form VISAWI-S [42] of the VISAWI questionnaire and clarity is rated with the new UEQ+ scale that was described in the previous section.

### A. Participants

A link to the online study was sent per mail to 8 classes of a vocational school for technology and design in Lingen (Germany). 168 subjects (135 males, 33 females, average age 22 years) participated in the study. Participation was voluntary and participants received no benefits for taking part in the study.

### B. Material

A fully functional booking portal for holiday trips with real content was used as stimulus.

To create some variety concerning aesthetic impression and clarity four layout variants were created. The CSS of the booking portal was manipulated to create a visually attractive (A), a visually unattractive (B), a version with a high level (C) and low level of clarity (D).

The booking portal was in addition manipulated in a way that the final confirmation step of a holiday booking does not really trigger the booking but navigates to pages that allow to rate the booking experience.

Some examples of pages in the booking portal are shown in Fig. 5 and Fig. 6.

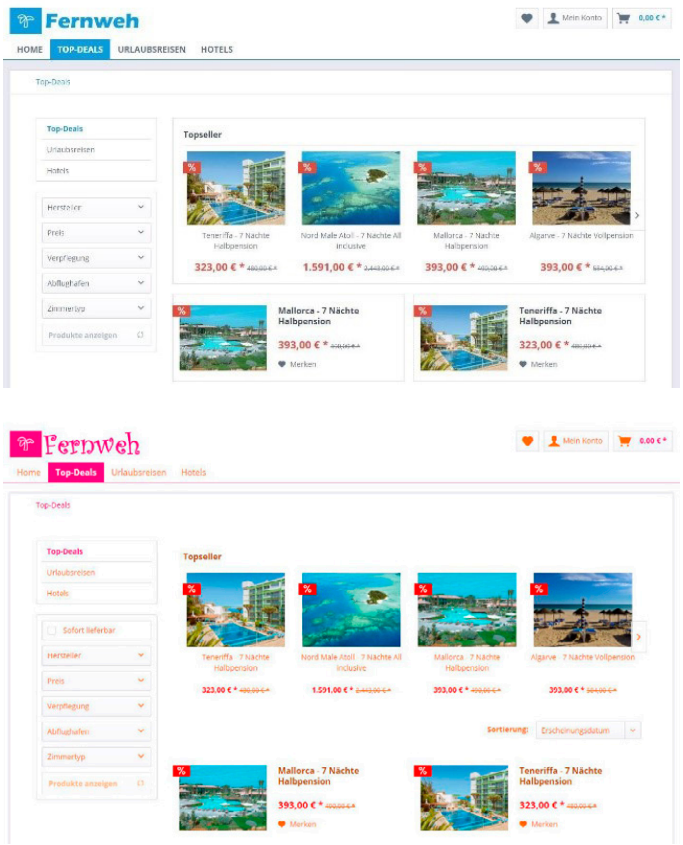



Fig. 5. Search result page for the visually appealing (top) and visually unappealing (bottom) condition. Manipulation of aesthetic appeal was done mainly by changing fonts and font respectively link colours.

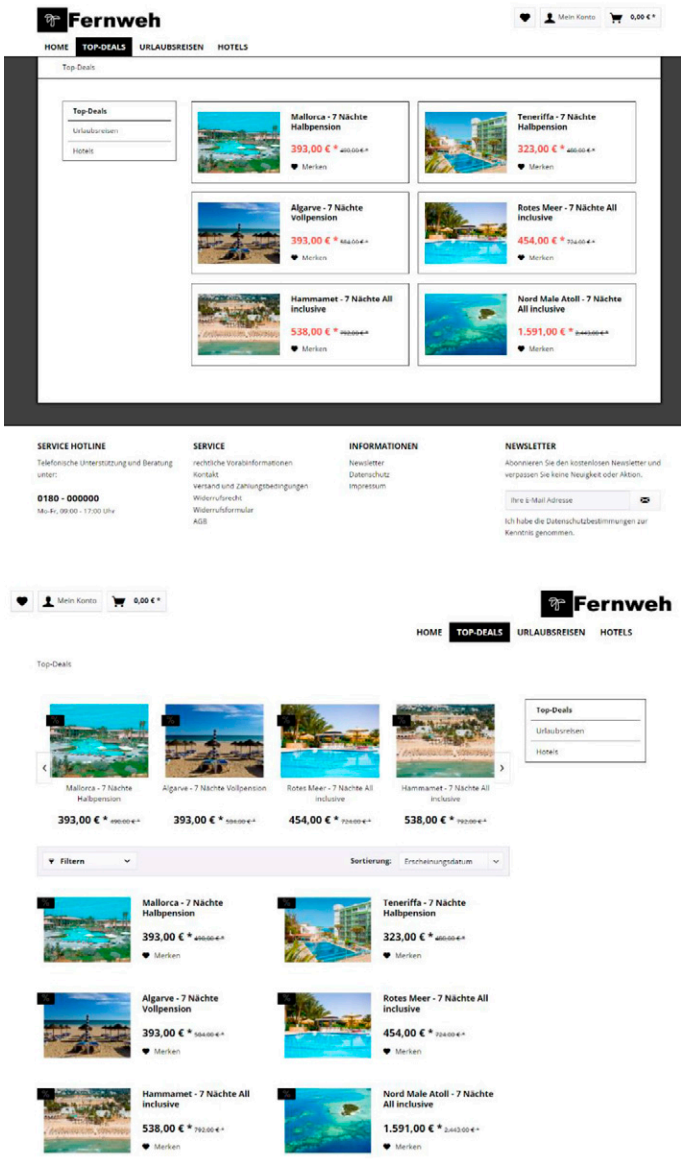



Fig. 6. Search result pages for the versions with high visual clarity (top) and low visual clarity (bottom). Manipulation was done by changing alignment, adding and removing elements and using structuring elements like boxes.

## C. Procedure

A link to the study was distributed per E-mail. When this link was clicked the participant was randomly assigned to one of the four layout variants. Each participant interacted only with one of these variants during the study.

On the start page of the study participants were instructed to the task. It was explained that they should book a holiday trip according to their personal preferences in a booking portal. They were informed that the confirmation step would not trigger a booking but navigate them to a questionnaire to rate the user experience of the portal.

At the bottom of this start page a link was placed that navigates to a page that asks for age and gender of the participant. From that page the booking portal could be started.

Inside the booking portal the navigation was not restricted. Participants could search for an interesting offer without limitations (all pages were accessible and there was no time limit).

After the participant has decided for a trip and clicked on the final booking step, he or she is redirected to a page that contains the four items of the short form of the VISAWI [42]. Once this has been filled in and the participant has submitted the answer, a page containing the 10 items of the SUS [41] is shown. Submitting the SUS data navigates to a page with the 4 items of the scale to measure clarity.

Once these data have been submitted a final page that allows some optional remarks or free text comments concerning the experiment is shown, and after this final page has been submitted, a page is shown that thanks for the participation.

## D. Results

Table I shows the mean scale values of the VISAWI-S, SUS and clarity scale for the four layout variants of the booking portal. This data shows that the intended manipulations of the layouts created the intended effect.

TABLE I. Mean Values of the Three Questionnaires Used to Measure Usability (SUS), Aesthetics (VISAWI-S) and Visual Clarity (New Scale). The VISAWI-S and Clarity Ratings Range From 1 (Worst) to 7 (Best), While SUS Ratings Range From 0 (Worst) to 100 (Best)

| Variant | VISAWI-S | SUS | Clarity |
|---|---|---|---|
| A (attractive) | 5.57 (1.09) | 80.51 (10.76) | 5.82 (0.99) |
| B (unattractive) | 3.24 (1.45) | 66.65 (15.19) | 4.20 (1.69) |
| C (high clarity) | 4.84 (1.17) | 82.62 (11.15) | 6.13 (0.79) |
| D (low clarity) | 3.85 (1.27) | 63.10 (17.19) | 3.67 (1.74) |

Now we concentrate on the mediator effect of visual clarity on the dependency of usability and aesthetics, which was the main goal of the replication study.

We first take a look at the correlations between the three variables over all three variants. The following highly significant correlations were observed:

- *Aesthetics, Usability*: r=0.679, (t(166)=11.91, p<0.001)
- *Aesthetics, Clarity*: r=0.715, (t(166)=13.18, p<0.001)
- *Usability, Clarity*: r=0.758, (t(166)=14.97, p<0.001)

The partial correlation between aesthetics and usability, if we control the impact of clarity on both variables, is reduced to 0.299 (t(168)=4.031, p < 0.01), which is again a first indicator for the assumed mediator effect.

We now describe the mediator analysis in detail in Fig. 7 and Fig. 8. The values can be interpreted as described above for Fig. 2.

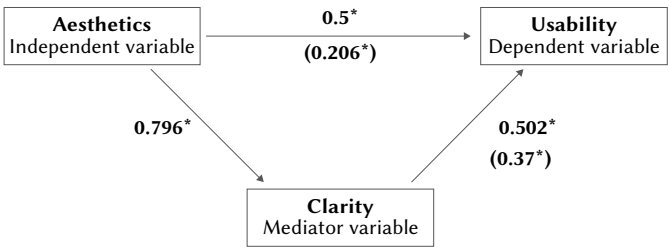



Fig. 7. Dependency between aesthetics and usability considering the impact of visual clarity. Regression coefficients all significantly >0, p<0.01).

Thus, again the values show that the influence of aesthetics on usability decreases if we consider clarity as a mediator variable. The Sobel test shows a significant mediator effect (Sobel z=6.878, p< 0.01).

For the opposite direction of the dependency the Sobel test shows again a significant mediator effect (Sobel z=5.465, p<0.01).

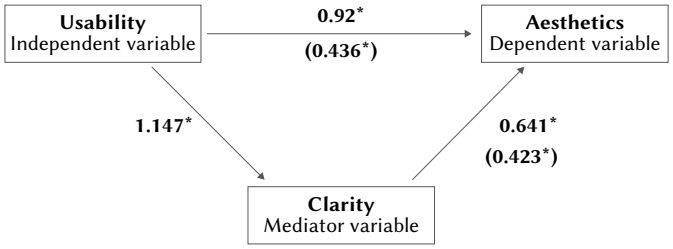



Fig. 8. Dependency between usability and aesthetics considering the impact of visual clarity. Regression coefficients all significantly >0, p<0.01).

### E. Conclusions

The mediator effect of study 1 could be reproduced. The effect is even a bit stronger (as can be seen by the reduction of the regression coefficients) than in the first study.

Thus, even if participants interact with the pages and if a different way to operationalise the three variables usability, aesthetics and visual clarity is chosen, the expected mediator effect is visible in the data.

## VII. Summary

Several different explanations have been proposed to explain the dependency between perceived usability and perceived aesthetics of a product. We suggest in this paper a new explanation that is based on the observation that items used by some UX questionnaires as an indicator for usability aspects are used in other questionnaires as an indicator for visual aesthetics. What is common to those items is that they describe the impression of clarity or visual simplicity of the layout.

We showed in two different studies that visual clarity acts as a mediator for the dependency between perceived usability and perceived aesthetics. This suggests that the impression of a user interface as clean, aligned, ordered and visually simple acts as a common factor that impacts aesthetics and usability ratings. This explanation allows to explain the dependency of usability and aesthetics in both directions (*What is beautiful is usable* and *What is usable is beautiful*) and is conceptually much simpler than other explanations.

Both studies had a quite different setup and the operationalisation of the variable's usability, aesthetics and clarity differed. Thus, the mediator effect could be detected under quite different settings for the study.

A practical advantage of this finding is that it is beneficial to invest a lot of effort in a visually clearly structured user interface during the design of new user interfaces. This will impact usability and aesthetic ratings. The good thing is that this aspect is not so difficult to handle from the point of view of a designer. Well-known design guidelines and heuristics, for example the minimisation of alignment lines in the layout, the number of different visual elements, the variety of elements sizes, etc. can be used to optimise a user interface under this aspect.

## References

[1] J. Preece, Y. Rogers, H. Sharp, *Interaction Design: Beyond Human-Computer Interaction*, Chichester, England: John Wiley and Sons Ltd, 2002.

[2] M. Hassenzahl, "The Effect of Perceived Hedonic Quality on Product Appealingness." *International Journal of Human-Computer Interaction 13(4)*, pp. 481-499, 2001, doi: 10.1207/S15327590IJHC1304_07.

[3] M. Kurosu, K. Kashimura, "Apparent usability vs. inherent usability: experimental analysis of the determinants of the apparent usability", *Conference Companion on Human Factors in Computing Systems*, Denver, USA, 1995, pp. 292-293, doi: 10.1145/223355.223680.

[4] N. Tractinsky, "Aesthetics and apparent usability: empirically assessing cultural and methodological issues.", Proceedings of the *ACM SIGCHI Conference on Human factors in computing systems*, New York, USA: Association for Computing Machinery, pp. 115-122, 1997, doi: 10.1145/258549.258626.

[5] N. Tractinsky, A.S. Katz, D. Ikar, "What is beautiful is usable." *Interacting with Computers, vol. 13*, pp. 127–145, 2000, doi: 10.1016/S0953-5438(00)00031-X.

[6] D. Norman, *Emotional Design: Why We Love (Or Hate) Everyday Things*, Boulder, USA: Basic Books, 2003.

[7] M. Thielsch, R. Haines, L. Flacke, *Experimental investigation on the effects of website aesthetics on user performance in different virtual tasks*, PeerJ 7:e6516, 2019, doi: 10.7717/peerj.6516.

[8] M. Thielsch, J. Scharfen, E. Masoudi, M. Reuter, "Visual aesthetics and performance: A first meta-analysis.", *Mensch und Computer 2019*, pp. 199-210, doi: 10.1145/3340764.3340794.

[9] W. Ilmberger, M. Schrepp, T. Held, "What kind of cognitive process causes the relationship between aesthetics and usability." in Holzinger, A. (ed.): USAB 2008, LNCS 5298, 2008, pp. 43-54, doi: 10.1007/978-3-540-89350-9_4.

[10] A. Tuch, S. Roth, K. Hornbaek, "Is beautiful really usable? Toward understanding the relation between usability, aesthetics, and affect in HCI.", *Computers in Human Behavior*, vol. 28, no. 5, pp. 1596-1607, 2012, doi: 10.1016/j.chb.2012.03.024.

[11] D. Winter, A. Hinderks, M. Schrepp, J. Thomaschewski, "Welche UX Faktoren sind für mein Produkt wichtig?", *Mensch und Computer 2017 - Usability Professionals*, Regensburg, Germany, 2017, doi: 10.18420/muc2017-up-0002.

[12] A. Hinerks, D. Winter, M. Schrepp, J. Thomaschewski, "Applicability of User Experience and Usability Questionnaires.", *Journal of Universal Computer Science*, vol. 25, no. 13, pp. 1717-1735, 2020, doi: 10.3217/jucs-025-13-1717.

[13] A. M. Isen, "Positive affect and decision making", in Lewis, M., Haviland, J.M. (ed.): Handbook of emotions (2nd edition), New York, USA: Guilford Press, pp. 417-435, 2000.

[14] N. Schwarz, "Situated cognition and the wisdom of feelings.", *The wisdom of feeling: Psychological processes in emotional intelligence*, New York, USA: Guilford Press, pp. 144–166, 2002.

[15] J. Kim, J. K. Moon, "Designing towards emotional usability in customer interfaces – trustworthiness of cyber-banking system interfaces.", in *Interacting with Computers*, vol. 10, pp. 1-29, 1998, doi: 10.1016/S0953-5438(97)00037-4.

[16] A. Rafaeli, I. Vilnai-Yavetz, "Instrumentality, aesthetics and symbolism of physical artefacts as triggers of emotion.", in *Theoretical Issues in Ergonomics Science*, vol. 5, pp. 91-112, 2004, doi: 10.1080/1463922031000086735.

[17] K. K. Dion, E. Berscheid, E. Walster, "What is beautiful is good.", in *Journal of Personality and Social Psychology*, vol. 24, pp. 285-290, 1972, doi: 10.1037/h0033731.

[18] A. Dick, C. Dipankar, B. Gabriel, "Memory-Based Inference During Consumer Choice.", in *Journal of Consumer Research*, vol. 17, pp. 82-93, 1990, doi: 10.1086/208539.

[19] G. T. Ford, R. A. Smith, "Inferential Beliefs in Consumer Evaluations: An Assessment of Alternative Processing Strategies.", in *Journal of Consumer Research*, vol. 14, pp. 363-371, 1987, doi: 10.1086/209119.

[20] C. E. Lance, J. A. LaPointe, A. M. Stewart, "A test of the context dependency of three causal models of halo rater error.", in *Journal of Applied Psychology*, vol. 79, no. 3, pp. 332-340, doi: 10.1037/0021-9010.79.3.332.

[21] T. Lavie, N. Tractinsky, "Assessing dimensions of perceived visual aesthetics of web sites.", in *International Journal of Human-Computer-Studies*, vol. 60, pp. 269-298, 2004, doi: 10.1016/j.ijhcs.2003.09.002.

[22] M. Thielsch, M. Mooshagen, "Erfassung visueller Ästhetik mit dem VISAWI.", *Usability Professionals 2011*, Stuttgart, Germany: German UPA e.V., pp.260-265, 2011.

[23] D. C. Ngo, L. S. Teo, J. G. Byrne, "Formalizing guidelines for the design of screen layouts.", in *Displays*, vol. 21, pp. 3-15, 2000, doi: 10.1016/S0141-9382(00)00026-3.

[24] G. A. Bonsiepe, "A method for quantifying order in typographic design.", in *Journal of Typographic Research*, vol. 2, pp. 203-220, 1968.

[25] L. Deng, M. S. Poole, "Aesthetic design of e-commerce web pages – Webpage Complexity, Order and preference.", in *Electronic Commerce Research and Applications*, vol 11, no. 4, pp. 420-440, 2012, doi: 10.1016/j.elerap.2012.06.004.

[26] R. Reber, N. Schwarz, P. Winkielman, "Processing Fluency and Aesthetic Pleasure: Is Beauty in the Perceiver's Processing Experience?", in *Personality and Social Psychology Review*, vol. 8, no. 4, pp. 364-382, 2004, doi: 10.1207/s15327957pspr0804_3.

[27] G. D. Birkhoff, *Aesthetic Measure*, Cambridge, USA: Harvard University Press, 1933.

[28] B. Laugwitz, T. Held, M. Schrepp, "Construction and evaluation of a user experience questionnaire.", in *Symposium of the Austrian HCI and Usability Engineering Group*, Herdelberg, Germany: Springer, pp. 63-76, 2008, doi: 10.1007/978-3-540-89350-9_6.

[29] M. Schrepp, A. Hinderks, J. Thomaschewski, "Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S).", in *IJIMAI*, vol. 4, no. 6, pp.103-108, 2017, doi: 10.9781/ijimai.2017.09.001.

[30] M. Schrepp, A. Hinderks, J. Thomaschewski, "Construction of a benchmark for the User Experience Questionnaire (UEQ).", in *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 4, pp. 40-44, 2017, doi: 10.9781/ijimai.2017.445.

[31] M. Hassenzahl, M. Burmester, F. Koller, "AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität.", in *Mensch & Computer 2003: Interaktion in Bewegung*, Stuttgart, Germany: B. G. Teubner, pp. 187-196, 2003.

[32] J. R. Lewis, "Psychometric evaluation of the post-study system usability questionnaire: The PSSUQ.", in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 36, no. 16, pp. 1259-1260, 1992, doi: 10.1177/154193129203601617.

[33] M. Thielsch, I. Blotenberg, R. Jaron, "User Evaluation of Websites: From First Impression to Recommendation", in *Interaction with Computers*, vol. 26, pp. 89-102, 2014, doi: 10.1093/iwc/iwt033.

[34] R. Rosenholtz, Y. Li, L. Nakano, "Measuring visual clutter.", in *Journal of Vision*, vol. 7, no. 2, pp.1-22, 2007, doi: 10.1167/7.2.17.

[35] K. Müller, M. Schrepp, "Visuelle Komplexität, Ästhetik und Usability von Benutzerschnittstellen.", *Mensch & Computer 2013 – Interaktive Vielfalt*, München, Germany: Oldenbourg Verlag, pp. 211-220, 2013, doi: 10.1524/9783486781229.211.

[36] T. K. Comber, J. R. Maktby, "Screen complexity and user design preferences in windows applications.", in *Harmony through working together: proceedings of OZCHI 94*, pp.133-137, 1994.

[37] D. C. Donderi, "Visual Complexity: A Review.", in *Psychological Bulletin 2006*, vol. 132, pp. 73-79, 2006, doi: 10.1037/0033-2909.132.1.73.

[38] M. Sobel, "Asymptotic confidence intervals for indirect effects in structural equation modelling.", in *Sociological Methodology*, vol. 13, pp. 290-312, 1982, doi: 10.2307/270723.

[39] M. Schrepp, J. Thomaschewski, "Design and Validation of a Framework for the Creation of User Experience Questionnaires.", in *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 7, pp. 88-95, 2019, doi: 10.9781/ijimai.2019.06.006.

[40] W. Revelle, psych: Procedures for Psychological, Psychometric, and Personality Researchm, Illionis, USA: Northwestern University, 2018, https://CRAN.R-project.org/package=psych Version = 1.8.12.

[41] J. Brooke, "SUS-A quick and dirty usability scale.", in Usability evaluation in industry, vol. 189, no. 194, pp. 4-7, 1996, doi: 10.1201/9781498710411-35.

[42] M. Moshagen, M. Thielsch, "A short version of the visual aesthetics of websites inventory", in *Behaviour & Information Technology*, vol. 32, no. 12, pp. 1305-1311, 2013, doi: 10.1080/0144929X.2012.694910.

Martin Schrepp

Martin Schrepp has been working as a user interface designer for SAP SE since 1994. He finished his diploma in mathematics in 1990 at the University of Heidelberg (Germany). In 1993 he received a PhD in Psychology (also from the University of Heidelberg). His research interests are the application of psychological theories to improve the design of software interfaces, the application of *Design for All* principles to increase accessibility of business software, measurement of usability and user experience, and the development of general data analysis methods. He has published several papers concerning these research fields.

Raphael Otten

Raphael Otten was born on Februar 19, 1990 in Lingen, Germany. He studied business informatics in the year 2011-2014 at the University of Applied Sciences Osnabrück, Germany. He graduated with a Bachelor of Science in the year 2014. After that, he studied media informatics in the years 2014-2019 at the University of Applied Sciences Emden/Leer. He graduated with a Master of Science in the year 2019. Since 2008, Raphael Otten is working as a Backend Developer at connectiv! eSolutions GmbH.

Kerstin Blum

Kerstin Eva Blum (formerly Mueller) was born on June 23rd, 1986 in Nürtingen, Germany. She studied psychology in the years 2007-2010 at the University of Konstanz, Germany. She graduated with a Bachelor of Science in the year 2010. After that, she studied psychology in the years 2010-2013 at the Ruprecht-Karls University of Heidelberg with a focus on "organizational behavior and adaptive cognition". She graduated with a Master of Science in the year 2013. Since 2013, Kerstin Blum is working as a User Experience Designer in various application areas at SAP SE.

Jörg Thomaschewski

Jörg Thomaschewski received a PhD in physics from the University of Bremen (Germany) in 1996. He became Full Professor at the University of Applied Sciences Emden/ Leer (Germany) in September 2000. His research interests are human-computer interaction, e-learning, and software engineering. Dr. Thomaschewski is the head of the research group "Agile Software Development and User Experience".

# DeepFair: Deep Learning for Improving Fairness in Recommender Systems

Jesús Bobadilla*, Raúl Lara-Cabrera*, Ángel González-Prieto, Fernando Ortega

ETSI Sistemas Informáticos, Universidad Politécnica de Madrid, Madrid (Spain)

Received 2 July 2020 | Accepted 10 November 2020 | Published 30 November 2020

## Abstract

The lack of bias management in Recommender Systems leads to minority groups receiving unfair recommendations. Moreover, the trade-off between equity and precision makes it difficult to obtain recommendations that meet both criteria. Here we propose a Deep Learning based Collaborative Filtering algorithm that provides recommendations with an optimum balance between fairness and accuracy. Furthermore, in the recommendation stage, this balance does not require an initial knowledge of the users' demographic information. The proposed architecture incorporates four abstraction levels: raw ratings and demographic information, minority indexes, accurate predictions, and fair recommendations. Last two levels use the classical Probabilistic Matrix Factorization (PMF) model to obtain users and items hidden factors, and a Multi-Layer Network (MLN) to combine those factors with a 'fairness' (ß) parameter. Several experiments have been conducted using two types of minority sets: gender and age. Experimental results show that it is possible to make fair recommendations without losing a significant proportion of accuracy.

## Keywords

Recommender Systems, Collaborative Filtering, Deep Learning, Fairness, Social Equality.

DOI: 10.9781/ijimai.2020.11.001

## I. Introduction

FAIRNESS in Recommender Systems (RS) is a very important issue, since it is part of the path to get a fair society. Nowadays, recommendations come to us from a variety of online services such as Netflix, Spotify, TripAdvisor, Facebook, Amazon, etc. All these services rely on hybrid RS [1] whose kernel is the Collaborative Filtering (CF). CF data is the set of the users' preferences on the items: tens or hundreds of millions of ratings, likes, clicks, etc. It seems great, since in theory, the more the data the better the recommendations; unfortunately, this data is usually biased [2]–[3] and minority groups are the most damaged ones. Common minority groups are female (vs. male) and senior (vs. young); both groups tend to receive unfair recommendations from online services. This situation has a perverse effect: a cycle that feeds back, where unfair recommendations make minority users to lose confidence in the system, to decrease their interaction and, thus, to receive even more unfair recommendations. The time has come to increase research in fair RS to reduce the digital gap [4]–[5] between minority and non-minority groups.

CF RS research has been traditionally focused in accuracy improvement [6], although some other objectives have increased the research attention in the last years: novelty [7], reliability [8], diversity [9] and serendipity [10]–[11] among them. Surprisingly, fairness has not been a main objective in the RS priorities. One of the reasons is the idea that improving fairness does not lead us to more valued recommendations, such as accuracy, novelty or diversity clearly do. Nevertheless, society needs to point in the opposite direction [12], and a set of new quality goals are growing [13]: relevance, fairness, and satisfaction among them. The historical development of CF has not helped to the fairness research, either: when the k-Nearest Neighbors (kNN) algorithm [14] dominated the field, it was less likely that a reduced set of neighbors produced biased recommendations. However, in a very short time the Matrix Factorization (MF) method prevailed as standard, and the fairness goal relevance grew up [15]. MF makes a compressed version of the ratings that belong to the dataset, catching the essence of them. The compressed models are sensible to the data biases such as the demographic ones: gender, age, etc. [16] making fairness a particularly relevant goal.

As a consequence of the CF research evolution, existing publications to improve fairness using the kNN algorithm are scarce; as an example, in [17] authors look for balanced neighborhoods as a mechanism to preserve personalization (accuracy) while enhancing the recommendations fairness. It is also remarkable the differentiation that takes place, in this context, between consumer-centered and provider-centered fairness. Fairness has been studied in the CF context in two main directions: a) finding that data biases really generates unfair recommendations, and b) providing quality measures or methods to quantify recommendations fairness. From the first block, in [18] authors argue that improving recommendations diversity leads to discrimination among the users and unfair results. The response of CF algorithms to the demographic distribution of ratings is studied in [19]; they find that common CF algorithms differ in the gender distribution of their recommendation lists. A preliminary experimental study on synthetic data was conducted in [20], where conditions under which a recommender exhibits bias disparity and the long-term effect of recommendations on data bias are investigated. From the second block (quality measures) in [21] they claim that biased data can lead CF

* Corresponding author.

E-mail addresses: jesus.bobadilla@upm.es (J. Bobadilla), raul.lara@upm.es (R. Lara-Cabrera).

methods to make unfair predictions for users from minority, and they propose new metrics that help reducing fairness. Disparity scores has also been proposed [18] to obtain fairness measures. Bias disparity can be defined as "how much an individual's recommendation list deviates from his or her original preferences in the training set" [20], whereas average disparity measures how much preference disparity between training data and recommendation list for the minority group of users is different from that for the non-minority group [22]. Fairness quality results in our paper implement these concepts.

Fairness in information retrieval has been focused on study data bias more than acting on the machine learning models: "teams typically look to their training datasets, not their machine learning models, as the most important place to intervene to improve fairness in their products" [12]. The machine learning achievements in the fairness issue have been reviewed in [23], where they find some "frontiers" that machine learning has not crossed yet. The MF disadvantages in CF have been studied in [21], where authors state that the MF model cannot manage the two main types of imbalanced data: population imbalance and observation bias. RS fairness has been even less covered in Deep Learning (DL) than in machine learning; as an example, in this current survey of RS based on DL [24] the fairness goal is not mentioned, not even in its "possible research directions" section. The same happens with the current review paper [25] where fairness is not mentioned despite the complete set of DL-based RS included in the publication. In fact, state of the art research in this area is focused on accuracy improvements [26]–[27] and it has not covered this subject. To afford a DL-based and fair RS is difficult due to the neural black box model [28], that is not easy to explain or vary. Nevertheless, to tackle CF fairness using DL has the advantage of providing a starting base where accuracy is high [29]; it is particularly convenient since the increase in fairness usually leads to the decrease in accuracy.

For the stated reasons, the hypothesis of this paper claims that it is possible to design a DL architecture that provides fair CF recommendations at the cost of reasonable decreases of accuracy. A DL approach to obtain fair recommendation provides a novel scenario in the RS field. This scenario opens the door to reach accurate and fair predictions, but it is not a straightforward how to make the architectural design: we have to deal not just with raw ratings data, but also with the necessary demographic information to determine the target minority groups: female vs. male, senior vs. young, etc. Moreover, the neural network learning model cannot be changed as easily as the kNN approach or even some machine learning algorithms. For all this, the proposed DL approach relies on an enriched set of input data and a tailored loss function that minimizes not only the accuracy errors but also the fairness ones. Fairness errors can be measured using the disparity scores concept [18], but how these scores are fed is a research open issue.

The proposed neural network learns from data that accomplish the current disparity concept: "deviation from the list of recommendations and the training data". We have specified it into two related indexes: the items one, that assigns a minority value to each item (e.g. a femininity value to a film, that depends on the female and the male preferences on this movie), and the users one, that assigns a minority value to each user (e.g. a femininity value to a user, that depends on the femininity of the items preferred for this user). Once both indexes have been set, it is possible to design a neural network loss function that rewards equality between each user minority value and his/her recommended items minority values. An additional design decision we have taken is to choose a regression approach [8] instead a classification one [27]: since we need to simultaneously minimize accuracy and fairness errors in the loss function, it is straightforward to pack them into a combined value so that the neural network provides us with balanced fairness/accuracy regression results. Finally, we have chosen a combined MF and DL approach [8] [30]; this design allows us to decouple the accuracy and the fairness abstraction levels by assigning accuracy to the MF and fairness to the DL stage.

A main advantage of the proposed architecture is that, once the model has learned, recommendations can be made to users that do not have associated demographic information; that is: we can fairly recommend to users without knowing its minority nature. It is possible because the neural network can learn the minority pattern in the same process that it learns to minimize the accuracy/fairness prediction error. It is a commercial advantage since many users avoid filling in their personal data.

In summary, designing recommender systems that are capable of providing fair recommendations without a high loss of accuracy is a significant contribution not only to the field of fairness in the ML-based RS, but also to the DL-based ones. As mentioned above, the former has merely proposed metrics for measuring unfairness in recommendations while the latter does not even consider fairness as a current goal.

As already discussed in Section I, existing recommender systems are primarily focused on providing recommendations as accurately as possible. Recommendations provided to minority groups of users are currently very unbalanced due to the RS datasets bias, and it leads to unfair recommendations made to the groups. State of the art in RS fairness is centered in memory-based methods, that are no longer commercially used due to their lack of accuracy. Research in model-based fair methods is scarce, and it is focused on trust-based systems, that usually require social information not available in most of the commercial RS. Our approach is a model-based one, making use of DL technology and which only needs the ratings information.

Based on the above, this paper's main research objective is to find a balance between accuracy and fairness in the recommendations made to the RS users. To this end, we propose a DL CF approach that can automatically adjust fairness and accuracy in recommendations.

The rest of the paper has been structured as follows: in Section II the proposed method is explained and the experiments design is defined. Section III shows the experiments' results and their discussions. Finally, Section IV contains the main conclusions of the paper and the future work.

## II. Materials and Methods

This section is devoted to describing our proposed method as well as the experimental setup we have used to evaluate it.

### A. Proposed Method

The proposed architecture incorporates four different abstraction levels, as depicted in Fig. 1, to get the desired fair recommendations: a) raw ratings and demographic information, b) minority indexes for both users and items, c) accurate predictions, and d) fair recommendations. Level 'b' just makes some simple statistical operations by combining ratings and demographic information; level 'c' uses the classical Probabilistic Matrix Factorization (PMF) model in order to obtain users and items hidden factors; finally, level 'd' makes use of a Multi-Layer Network (MLN) to combine hidden factors and a 'fairness' (ß) parameter. This MLN generates the desired fair recommendations.

We will develop each of the three levels that make up our architecture: first, in the lowest level we create two related indexes: 1) *items minority index* (*IM*), and 2) *users minority index* (*UM*). The *IM* index will assign a minority value to each item in the dataset, e.g. when the minority group is 'female' we could call to the index 'femininity'. It will contain values [-1, 1] where negative ones mean feminine preferences and positive ones mean masculine preferences.

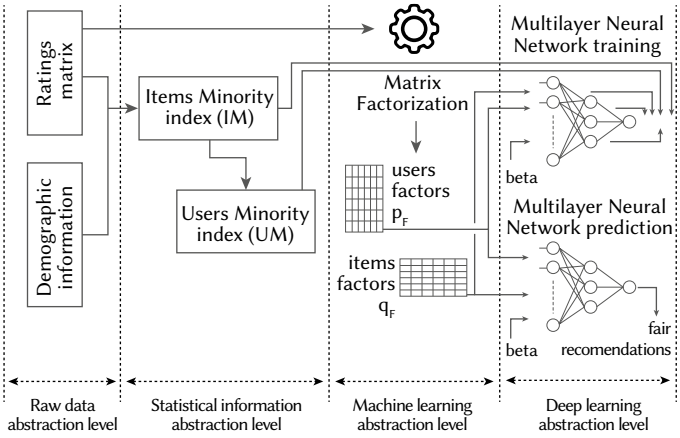



Fig. 1. Architecture overview.

Then, when an item has been assigned a negative value it means that it has been rated better by women than men. Once the *IM* index has been created it contains the minority values of all the items. By using the *IM* index, we will create the *UM* index. The *UM* index will assign a minority value to each user in the dataset. It also will contain values [-1, 1], where negative ones mean minority preferences and positive ones mean not minority preferences (masculine, in our example). A user assigned a negative *UM* value means that this user prefers negative *IM* items, and vice versa. Please note that, on many occasions, female users may have assigned positive *UM* values and male users may have assigned negative *UM* values, since there exist women with masculine preferences and men with feminine ones; same as young and older persons or any other minority versus majority groups. Thus, an important concept is that both the *IM* and *UM* indexes do not contain disjoint minority/majority demographic values; they contain minority/majority preferences. This design accurately fits the existing diversity of preferences contained in the CF based RS.

Now, we will explain the *IM* and *UM* indexes design that we will take as a base to get fair recommendations in the DL stage. First, we will differentiate between relevant and not relevant votes: relevant votes are those that indicate that the user liked the item; conversely not relevant votes (in our context) are those that indicate that the user did not liked the item. There can also exist votes that indicate indifference on the part of the user. In our formulation, relevant and not relevant votes are chosen by means of two thresholds; e.g. in a dataset where votes must be in the set {1, 2, 3, 4, 5} we can establish 4 as the relevant threshold and 2 as the non-relevant threshold. In this way the relevant set is {5, 4}, the non-relevant set is {2, 1} and {3} would be the 'indifference' set.

We define the *IM* index (11) for each item *i* as the majority score of *i* minus the minority score of *i*. The majority score (resp. minority score) of the item *i* is the number of majority (resp. minority) users that voted *i* as relevant minus the number of majority (resp. minority) users that voted *i* as non-relevant, divided by the total amount of majority (resp. minority) users that did not consider *i* as indifferent, see Equations (9) and (10) (resp. (7), (8)). When the proportion of the minority user preferences exceeds the proportion of the non-minority ones, the *IM* index values are negative. In the gender example, equation (11) can be read as: "proportion of males that liked item *i* minus males that did not like it, minus the proportion of females that liked item *i* minus females that did not like it". We have also set a minimum number of 5 votes to consider both the minority and non-minority sides of equation (11).

Once the *IM* index has been created, we can use it to establish the *UM* index values. Each *UM* value corresponds to a user of the RS dataset, and it provides the minority value of the user. Each user minority value will be defined by the minority of his/her preferences: to obtain each user *UM* value we just make the average of the *IM* minority values of the items that the user has voted, weighting each *IM* minority value with its corresponding user rating. Equation (13) models the explained behavior.

Let $\Theta\uparrow$ be the like threshold (1)

Let $\Theta\downarrow$ be the dislike threshold (2)

Let I be the set of items in the dataset (3)

Let U be the set of users in the dataset (4)

We will assign the following meanings to super index numbers: *m* for minority and *M* for non-minority:

Let $U^m$ be the set of minority users (5)

Let $U^M$ be the set of non-minority users (6)

Let $U_\uparrow(i)=\{u \in U | r_{u,i} \geq \Theta_\uparrow\}$ be the set of users who liked item i (7)

Let $U_\downarrow(i) = \{u \in U | r_{u,i} \leq \Theta_\downarrow\}$ be the set of users who did not like item *i* (8)

The majority score is

$$\mathfrak{I}^{\mathfrak{M}}(i) = \frac{|U_\uparrow(i) \cap U^M| - |U_\downarrow(i) \cap U^M|}{|U_\uparrow(i) \cap U^M| + |U_\downarrow(i) \cap U^M|} \quad (9)$$

The minority score is

$$\mathfrak{I}^{\mathfrak{m}}(i) = \frac{|U_\uparrow(i) \cap U^m| - |U_\downarrow(i) \cap U^m|}{|U_\uparrow(i) \cap U^m| + |U_\downarrow(i) \cap U^m|} \quad (10)$$

The *IM* and *UM* indexes are

$$IM(i) = \mathfrak{I}^{\mathfrak{M}} - \mathfrak{I}^{\mathfrak{m}} \quad (11)$$

$$IM = \{(i, IM(i)) | i \in I\} \quad (12)$$

$$UM(u) = \frac{\sum_{\{i \in I | r_{u,i} \neq \circ\}} \left(r_{u,i} - \frac{\Theta_1 + \Theta_2}{2} \cdot IM(i)\right)}{\left(N - \frac{\Theta_1 + \Theta_2}{2}\right) \cdot |\{i \in I | r_{u,i} \neq \circ\}|} \quad (13)$$

$$UM = \{(u, UM(u)) | u \in U\} \quad (14)$$

where $\circ$ means "not voted item" and *N* is the maximum possible vote.

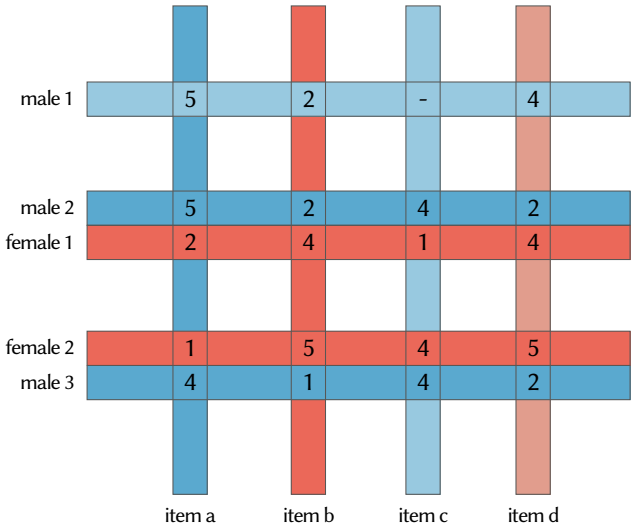



Fig. 2. Data-toy example to get *IM* and *UM* minority values.

Fig. 2 shows a data-toy example containing five users and four items. We will suppose that women are a minority group in this RS, compared to the men. We can observe that '*item a*' is clearly 'masculine', since it has been voted as 'relevant' for all the male users and it has been voted as 'non-relevant' for all the female users. The opposite situation is stated in '*item b*': it is a 'feminine' item according to the female relevant votes and the male non-relevant ones. '*Item c*' is quite masculine, although a female user liked it. Finally, '*item d*' shows the opposite situation to '*item c*'. According to it, the proposed *IM* equations return the following item minority values:

{⟨item a, 1⟩,⟨item b, −1⟩,⟨item c, 0.5⟩,⟨item d, −0.6⟩}

that fits with the explained behavior (Table I). Once the items' minority values *IM* are obtained, we can get the users minority ones (*UM*). First, we can observe how '*male 2*' and '*male 3*' users in the data-toy example have casted very 'masculine' ratings, since they have voted 'relevant' to the more 'masculine' items, and 'non-relevant' to the more 'feminine' items. This is not the case for the '*male 1*' user, that has a 'relevant' vote casted on the 'feminine' '*item d*'. The female users comparative is more complicated: '*female 1*' has casted all her votes in a 'feminine' way, whereas the '*female 2*' vote to the 'masculine' '*item c*' was 'relevant'; nevertheless, the '*female 2*' feminine votes are higher than the '*femenine 1*' ones. In this way, we expect the following results: a) positive *UM* values to male users and negative ones to female users, and b) a more 'minority' (feminine) value be assigned to '*male 1*' than to '*male 2*' and '*male 3*' . Table I shows the Fig. 2 data-toy *IM* results and Table II shows the *UM* ones.

TABLE I. DATA-TOY IM RESULTS

| Item | Value |
|---|---|
| **a** | [(3−0)−(0−2)]/5 = 1 |
| **b** | [(0 − 3) − (2 − 0)]/5 = −1 |
| **c** | [(2−0)−(1−1)]/4 = 0.5 |
| **d** | [(1 − 2) − (2 − 0)]/5 = −0.6 |

TABLE II. DATA-TOY UM RESULTS

| Item | Value |
|---|---|
| **male 1** | (5−3)·1+(2−3)·(−1)+(4−3)·(−0.6)=2.4/5=0.48 |
| **male 2** | (5−3)·1+(2−3)·(−1)+(4−3)·0.5+(2−3)· (−0.6)=4.1/5=0.82 |
| **female 1** | (2−3)·1+(4−3)·(−1)+(1−3)·0.5+(4−3)·(−0.6)= −3.6/5=−0.72 |
| **female 2** | (1−3)·1+(5−3)·(−1)+(4−3)·0.5+(5−3)·(−0.6)= −4.7/5= −0.94 |
| **male 3** | (4−3)·1+(1−3)·(−1)+(4−3)·0.5+(2−3)·(−0.6)= 4.1/5=0.82 |

Our architecture uses the PMF method to reduce the ratings matrix dimension and to get a condensed knowledge representation. From the condensed results we will be able to make accurate predictions. Equations (15)-(24) show the model formalization: the original ratings matrix is condensed in the two lower dimension matrices $P$ and $Q$ (equation (15)). $P$ is the users' matrix and $Q$ is the items' matrix. Both $P$ and $Q$ have a common dimension of $F$ hidden factors, where $F \ll M$ and $F \ll N$ (note that $M$ is numbers of users, and $N$ the number of items). Once the model has learnt, each user will be represented by a vector $\overrightarrow{p_u}$ of F factors, and each item will be also represented by a vector $\overrightarrow{q_i}$ of F factors. Each prediction of an item $u$ to a user $i$ is obtained by processing the dot product of these vectors (equation (16)). Since the users and the items hidden factors share the same semantic, predictions will be relevant when high values (positive or negative) of the factors line up in each user and item.

$$R \approx \hat{R} = P \cdot Q^t \tag{15}$$

$$\widehat{r_{u,i}} = \overrightarrow{p_u} \cdot \overrightarrow{q_i} = \sum_{f=1}^{F} p_{u,f} \cdot q_{i,f} \tag{16}$$

The $P$ and $Q$ factors will be used in our architecture to feed the DL process input as well as to set the output target labels. Factors are obtained by means of the gradient descent algorithm. The loss function just minimizes the prediction error: the difference between the predicted value and the existing rating (equation (17)).

$$\text{loss}(u,i) = \left(r_{u,i} - \widehat{r_{u,i}}\right)^2 \tag{17}$$

In order to achieve the gradient descent minimization process we obtain the partial loss derivatives: $\delta\text{loss}/\delta\overrightarrow{p_u}$ and $\delta\text{loss}/\delta\overrightarrow{q_i}$ (equations (18) and (19)).

$$\frac{\delta\text{loss}}{\delta\overrightarrow{p_u}} = \frac{\delta}{\delta\overrightarrow{p_u}}\left(r_{u,i} - \overrightarrow{p_u} \cdot \overrightarrow{q_i}\right)^2 = -2\overrightarrow{q_i} \cdot \left(r_{u,i} - \overrightarrow{p_u} \cdot \overrightarrow{q_i}\right) = -2\overrightarrow{q_i} \cdot e_{u,i} \tag{18}$$

$$\frac{\delta\text{loss}}{\delta\overrightarrow{q_i}} = \frac{\delta}{\delta\overrightarrow{q_i}}\left(r_{u,i} - \overrightarrow{p_u} \cdot \overrightarrow{q_i}\right)^2 = -2\overrightarrow{p_u} \cdot \left(r_{u,i} - \overrightarrow{p_u} \cdot \overrightarrow{q_i}\right) = -2\overrightarrow{p_u} \cdot e_{u,i} \tag{19}$$

This gives rise to the corresponding gradient descent factors update Equations (20) and (21).

$$p'_{u,f} = p_{u,f} + 2\gamma \cdot q_{f,i} \cdot e_{u,i} \tag{20}$$

$$q'_{f,i} = q_{f,i} + 2\gamma \cdot p_{q,f} \cdot e_{u,i} \tag{21}$$

Finally, we can add a regularization term for controlling the growing of the factors during the learning process, which gives rise to the loss function and the update rules shown in Equations (22) to (24).

$$\text{loss}(u,i) = \left(r_{u,i} - \widehat{r_{u,i}}\right)^2 + \frac{\lambda}{2}\sum_{f=1}^{F}(|P^2| + |Q^2|) \tag{22}$$

$$p'_{u,f} = p_{u,f} + \gamma\left(2q_{f,i} \cdot e_{u,i} - \lambda \cdot p_{uf}\right) \tag{23}$$

$$q'_{f,i} = q_{f,i} + \gamma\left(2p_{uf} \cdot e_{u,i} - \lambda \cdot q_{f,i}\right) \tag{24}$$

The highest semantic level of the proposed architecture is based on an MLN. Our MLN (see Fig. 3) model will take input vectors containing the following information: a) user hidden factors $p_u$, b) item hidden factors $q_i$, and c) $\beta \in [0,1]$ value. The β parameter is used to balance fairness and accuracy in predictions and recommendations: high β values will enhance accuracy, whereas low β values will enhance fairness. This balance is a key objective of our method: "To obtain fair recommendations just losing an acceptable degree of accuracy". Please note that we do not include demographic information to feed the MLN input, so once the MLN has learnt it will be able to make fair recommendations to users that have not filled demographic forms asking for gender, age, etc. This is an important commercial advantage, since it allows to make better marketing processes, to improve fairness, to focus prediction tasks, etc. It is also a challenge to the proposed machine learning framework because it is more difficult to increase recommendation fairness when demographic data is missing. The learning process has been based on input vectors containing the specified three information sources: : $\langle p_{u,f}, q_{f,i}, \beta\rangle$. We have set 11 input vectors to the MLN for each (user u, item i) rating of the dataset:

$$\langle p_{u,f}, q_{f,i}, 0.0\rangle, \langle p_{u,f}, q_{f,i}, 0.2\rangle, \ldots, \langle p_{u,f}, q_{f,i}, 1.0\rangle$$

The objective is to teach to the neural network on eleven fairness levels for each rating, as it can be seen in the left side of Fig. 3.

Once the MLN input vectors have been established, it is necessary to define their corresponding output labels to let the back-propagation algorithm learn the pattern. In our case we will design a loss function that minimizes both the prediction error and the fairness error. Equation (25) shows the typical prediction loss function, as we did in

equation (17). We define the fairness error as the distance between the user's minority and the item's minority; e.g. films recommended to a user (male or female) with an assigned 0.8 *UM* femininity value should be as similar as possible to a 0.8 IM in order to fit in the fairness issue. Since *UM* and *IM* vector values do not have the same distribution, we will apply a [0, 1] normalization in both of them and we will use the *UM'* and *IM'* names for the normalized versions. Then, to obtain the fairness error we establish equation (26). Finally, to combine equation (25) (accuracy) and equation (26) (fairness) the β parameter is added (equation (27)).

$$e_{u,i}^{accuracy} = \left(r_{u,i} - \sum_{f=0}^{F} p_{u,f} \cdot q_{i,f}\right)^2 \quad (25)$$

$$e_{u,i}^{\text{fairness}} = (IM_i' - UM_u')^2 \quad (26)$$

$$\text{loss}_{\text{u,i}} = \beta \cdot e_{u,i}^{\text{accuracy}} + (1-\beta) \cdot e_{u,i}^{\text{fairness}} \quad (27)$$

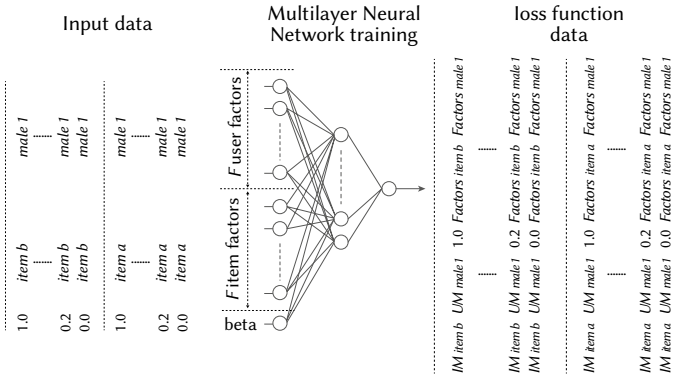



Fig. 3. Training information for the proposed MLN.

In the feed forward prediction stage, for each testing input data $\langle p_{u,f}, q_{f,i}, \beta \rangle$, the proposed neural network returns a real number whose meaning is the predicted loss error for the item $i$ to the user $u$ recommendation. The lower the predicted loss error, the better the combined ⟨accuracy, fairness⟩ values given the chosen ß accuracy vs. fairness balance. Once the network has learnt and the RS is in production phase, to make recommendations to an active user $u$, first we fix the ß value and then we feed the MLN with all the inputs $\langle \vec{p}_u, \vec{q}_i, \beta \rangle$ where $i$ runs over the set of items that the user $u$ has not voted (equation (28)).

$$X = \{\langle p_{u,f}, q_{f,i}, \beta \rangle | u \in U, i \in I, r_{u,i} \neq \circ\} \quad (28)$$

The set of N recommendations for the user $u$, $Z_{u,N}$ is the collection of N items with minimum loss function $h(\vec{p}_u, \vec{q}_i, \beta)$, where the $h$ function represents N feed forward operations.

### B. Experimental Setup

Experiments have been conducted using a well-known dataset called MovieLens 1M [31]. It contains 1,209,000 votes, 6040 users and 3952 items. We have used eleven different values of the β parameter (from 0.0 to 1.0, step 0.2); consequently, the MLN has been trained using 13,299,000 input vectors and output target values. Training, validation, and test sets have been established: 70%, 10% and 20%, respectively. The PMF process has been run using 30 hidden factors (*F*), 80% training ratings, 20% testing ratings. Please note that these are the MLN parameters of the proposed method, different to the previously ones specified for the DL stage. The designed MLN contains an input layer of 30+30+1 = 61 values (Fig. 3). The first MLN internal layer has been set to 80 neurons (*relu* activation), followed by a 0.2 dropout layer to avoid overfitting. The second internal layer has been set to 10 neurons (*relu* activation) and, finally, the output layer contains just one neuron with no activation function. The chosen loss function has been *mae* and the optimizer *rmsprop*.

## III. Results

The experiments we have conducted are:

- Item Minority Index (IM) and User Minority Index (UM) distributions.
- User Minority Index (*UM*) comparative between each minority and non-minority group.
- Fairness prediction improvement using the heuristic algorithm.
- Fairness recommendation improvement using the heuristic algorithm.
- Fairness error and accuracy error for recommendations using the proposed DL architecture.

This section contains a subsection for each of the above set of performed experiments. We have selected two types of minority sets: a) gender: female vs. male, and b) youth: young vs. senior. Results are provided showing both minority types in two separated graphs of each figure. The MovieLens dataset, like in many other CF RS happens, is biased towards male and young people.

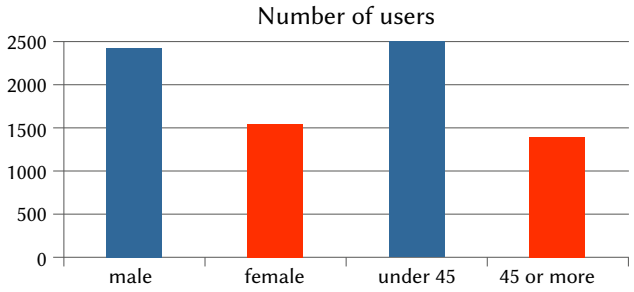



Fig. 4. Proportion of users in the MovieLens gender and age minority and non-minority groups.

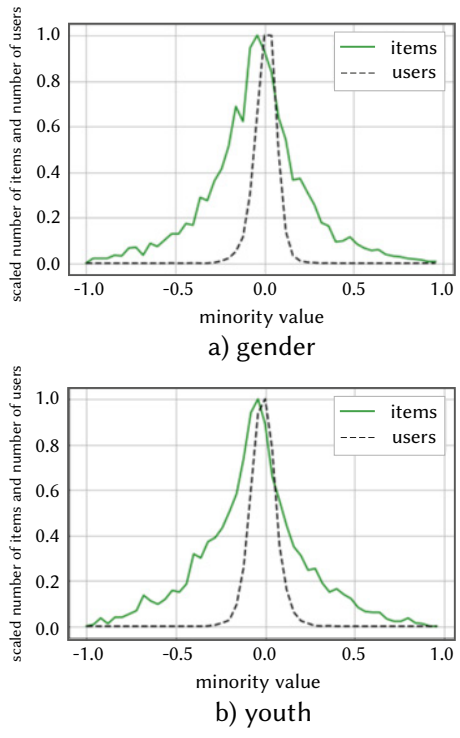



Fig. 5. Item Minority Index (*IM*) and User Minority Index (*UM*) distributions.

Thus, the chosen minority types are relevant and representative for this experimental study. Specifically, the MovieLens dataset contains more males than females; most of them are under 45 years old. Fig. 4 shows the proportions. Equations (11) and (13) describe both indexes behavior. The IM index semantic is simple and convincing, but it is necessary to be aware that we are not working with absolute values: in order to prevent data biases and to maintain the index values in a bounded range, we are working with preferences proportions; e.g. "proportion of male users that liked the items minus proportion of female users that liked the item". Since we expect a significant number of items that both minority and non-minority groups simultaneously like or dislike, *IM* proportions will be similar for both groups and consequently a significant number of *IM* values will concentrate around the 0.0 value. Fig. 5 shows the items and users minority indexes distributions, both for the gender and the youth minority groups.

The *UM* index values are obtained from the ratings that each user has casted to the items and from the *IM* value of each of those items. We can see in Fig. 5 that the users *UM* indexes (both for gender and youth) have a large concentration of values around 0. It provides us an important conclusion:

"In the reference dataset, most users have similar preferences regarding to the chosen minority groups". Looking at the *UM* distributions we can also yield another main conclusion: "Although users have similar preferences, there is a clear separation between minority groups" (left and right side of the graphs). Since the *UM* index is only used to feed internal DL processes the relevant information here is the proportion of the differences between values, and not their absolute values.

### *A. User Minority Index (UM) Comparative Between Each Minority and Non-minority Group*

In the above section we have confirmed two facts: 1) Users preferences are similar, even if they belong to different minority groups, and 2) Despite the previous conclusion, there is room to find minority behaviors of users. In this section we deepen in the minority *UM* values of users, to clear out our specific groups: male vs. female and senior vs. young. Fig. 6 shows the results: we can observe, in both cases, that groups have different behaviors and that they share a relevant number of preferences. Groups present different behaviors because they do not completely intersect their user minority values; as expected, minority groups return a mean less than zero whereas non-minority groups return it greater than zero. Groups share a relevant number of preferences because there exist a proportion of minority and non-minority users that share *UM* values (areas around 0.0 under both curves).

TABLE III. Users Classification Attending to the Minority/non-minority Groups

| group | type | correct | incorrect | correct % |
|---|---|---|---|---|
| **gender** | female | 1147 | 562 | 67.11 |
| | male | 3648 | 683 | 84.22 |
| **youth** | senior | 1231 | 195 | 86.32 |
| | young | 3144 | 1470 | 68.14 |

Due to the explained results, we can confirm that there is a not negligible proportion of minority users with non-minority preferences and vice versa. In any case, it varies depending on the specific minority group. As an example, we can observe in Fig. 6 how senior users have much less non-minority preferences than female ones, since there are small amounts of senior users whose minority value is greater than zero. Results show the convenience of using modern machine learning approaches to make fair recommendations to those users that share minority and non-minority preferences. Table III shows the specific number of users that have been classified as belonging to the minority or to the non-minority groups. Minority users (female, young) have an expected *UM* index less than zero. Non-minority users (male, senior) have an expected *UM* index greater than zero.

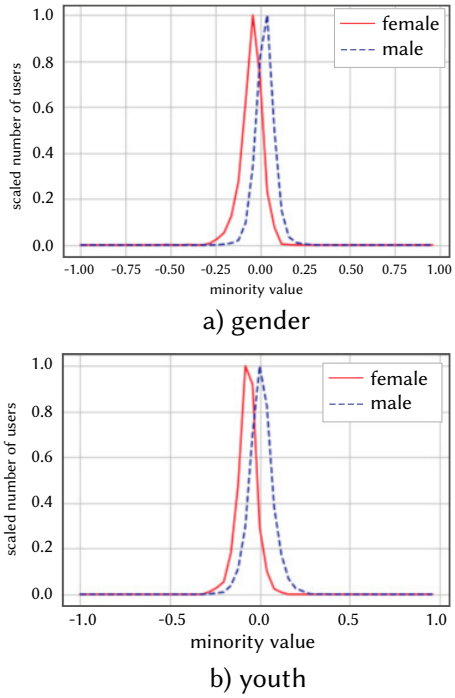



Fig. 6. User Minority Index (*UM*) comparative.

### *B. Fairness Prediction Improvement Using a Heuristic Algorithm*

Fig. 6 and table IV show us that most of the users are correctly grouped attending to their *UM* indexes, especially for seniors and males. They also show a considerable number of cases incorrectly classified, particularly for young and female groups. In this situation, we will obtain predictions from the test set and then check their quality in terms of the *IM* index. Table IV contains these experiments results: the IM averages fit the expected ranges (negative *IM* average for minority users, and positive *IM* average for non-minority users). Despite these positive results, ranges can be too narrow to ensure fair predictions. On the other hand, there will be situations in which it is intended to force the recommendations of an RS to move towards minority items, or perhaps towards majority items, depending on the type of users and/or the company policy.

TABLE IV. Averaged IM Values for the Predictions Made to Each Users' Group

| | female | male | senior | young |
|---|---|---|---|---|
| **IM mean** | -0.014 | 0.041 | -0.025 | 0.028 |

By filtering on the *IM* index, we can discard those predictions greater than a negative threshold and, in this way, increase the proportion of minority predictions. In the same way we can filter those predictions less than a positive threshold to increase the proportion of majority predictions. We have performed this experiment, calling alpha to the threshold. We can observe the expected behavior in Fig. 7, where growing minority (and majority) *IM* values are obtained in predictions when the alpha parameter increases. It also can be seen that the non-minority users (male, young) always obtain better predictions due to the RS datasets biases. Finally, we can state that, in this case, minority values can reach the starting majority ones by using low values of the alpha parameter (0.025 for gender and 0.05 for age).

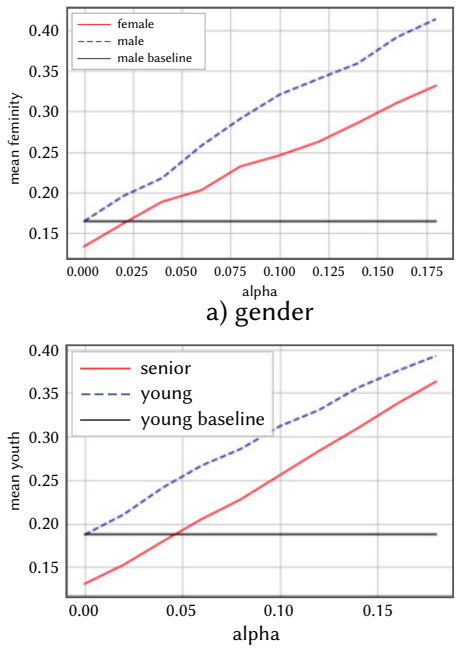



Fig. 7. Groups quality improvement by filtering predictions. x axis: alpha values used to filter on the IM items index. y axis: averaged minority of the filtered predictions. Minority (female, senior) curves are drawn using their absolute values.

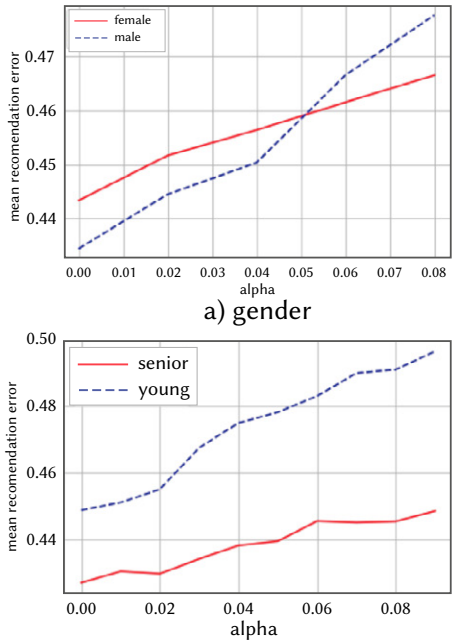



Fig. 8. Recommendation quality obtained by filtering predictions. x axis: alpha values used to filter on the IM items index. y axis: averaged error of the N recommendations. Lower error values are the better ones.

### C. Fairness Recommendation Improvement Using the Heuristic Algorithm

The previous section results show that it is possible to provide a heuristic method to improve recommendations fairness. To conduct the experiment, from the alpha filtered predictions (Fig. 7), we extract the N ones that provide higher prediction values, as usual in the CF operation. Thus, the complete recommendation method involves three sequential phases: 1) to obtain all the prediction value, minority value pairs, 2) to filter the pairs according to the minority threshold alpha parameter and each *minority value*, and 3) to select the N filtered predictions that have the N highest *prediction value* values.

Results in Fig. 8 show the existing correlation between recommendation errors and each chosen alpha value: the highest the alpha value, the better the recommendations fairness (Fig. 7), but as expected, also the worst the recommendation accuracy (higher error values in Fig. 8). Of course, we pay an accuracy price when we force fairer recommendations.

We have chosen a value of N = 10 recommendations to process the set of experiments. From Fig. 7 it can be observed that in the `youth' experiment our method provides better results (lower errors) for the minority `senior' group than for the `young' one. This is a good indication of the proposed heuristic method functioning. The `gender' experiments provide improvement in the minority female group from a specific value threshold (alpha = 0:05). All these results are consistent with Tables II and III values.

### D. Fairness Error and Accuracy Error for Recommendations Using the Proposed DL Architecture

Results obtained in the previous subsection tell us that we have designed a method that correctly provides fair recommendations. It is a simple, functional, and easy to implement machine learning approach. Nevertheless, it has some drawbacks:

- Choosing the adequate parameter alpha requires a fine-tuning process.
- Since the parameter alpha sign (less than or greater than zero) depends on the minority or non-minority nature of the recommended user, this recommendation method can only be applied to users with associated demographic information.

This subsection provides a DL approach that works without the above drawbacks. This method only needs the parameter $\beta$: it is used to select the accuracy vs. fairness balance. The $\beta$ range is [0, 1], whether 0 means 100% fairness and 0% accuracy, and 1 means 100% accuracy and 0% fairness. As it can be seen, to choose a $\beta$ value is straightforward and intuitive. Moreover: the chosen $\beta$ value does not change when the user is a minority one or he is not.

The proposed DL recommendation method explained in section 2 returns the results shown in Fig. 9. Graphs on the left of the figure contain the main information. Graphs on the right are [0, 1] scaled to find the optimum accuracy vs. fairness balances. The averaged error of the recommendations (equation (25)) is plotted using black lines. Dotted and dashed lines show the minority errors (equation (26)); that is: the distance between the minority value of each recommended user (UM) and the average of the minority values (IM) of their N recommended items. We are looking for recommended items in the minority range of the user; e.g. if a user (male or female) has an $UM$ = 0.7 (quite masculine), recommended items near $IM$ = 0.7 are the fairest ones, and they generate a low minority ('femininity') error.

'Gender' results are shown in the top-left graph of Fig. 9: as expected, accuracy increases (error decreases) as β increases (more importance to accuracy). The price to pay for this accuracy improvement is the simultaneous increase in the fairness error values. As β decreases (more importance to fairness), the opposite happens: higher prediction errors and lower fairness errors. 'Youth' results are shown in the low-left graph of Fig. 9: curve trends are like the 'gender' results. Graphs on the right of Fig.9 show the same results by using a normalized y axis: in this way we can find the optimum β values to balance accuracy and fairness in the recommendation task. To optimize results in this experiment, it is necessary to choose a β = 0.4 value: a balanced selection, something scored to the fairness objective. This result tells us that the balanced option (β = 0.5) can be the default one.

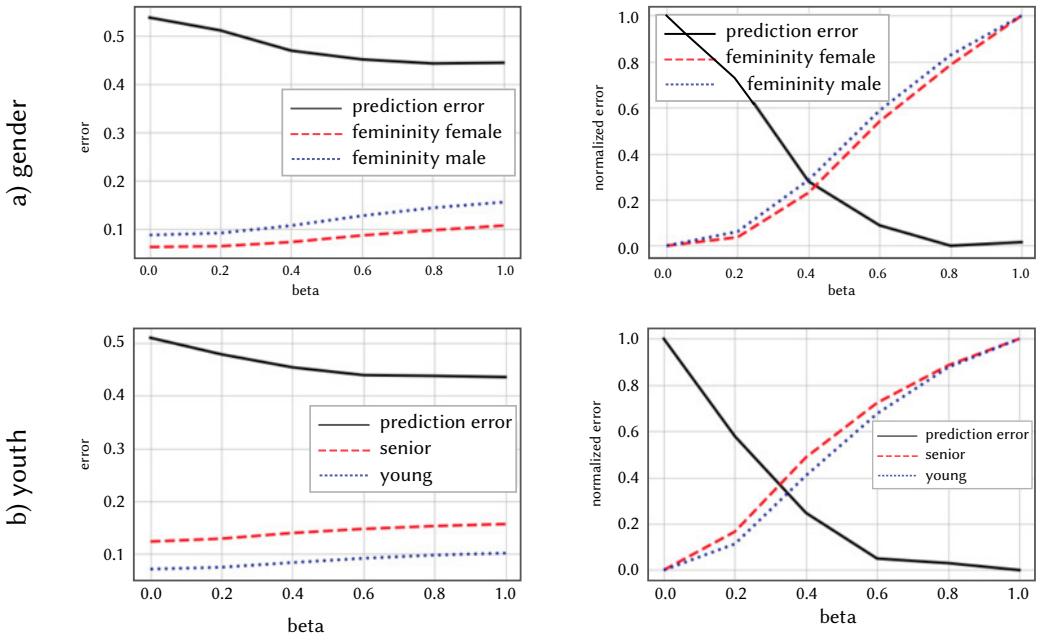



Fig. 9. Recommendation results using the proposed DL approach. y axis: averaged N = 10 recommendations error (normalized in the right graphs); x axis: β balance between fairness and accuracy (0.0 means 100% fairness and 0% accuracy, and 1.0 means 100% accuracy and 0% fairness).

## IV. Conclusions

Attending to the obtained results, it is understood that designing methods to improve CF fairness is not a simple task, but it is possible to take it out. Due to the fact that an appreciable proportion of minority and non-minority users share preferences it is necessary to make use of modern machine learning approaches in order to make fair recommendations not only to the 'purest' minority or non-minority users, but also to the users that mix some proportion of minority and non-minority preferences.

State of the art shows a lack of DL approaches to tackle fairness in RS, probably due to the neural networks black box model. The proposed method in this paper relies on an original loss function and input data to balance fairness and accuracy. This method combines several abstraction levels, and it can serve as baseline to DL future works in the field. An original architecture is provided, where machine learning and DL models are combined to obtain balanced accuracy vs. fairness recommendations. The architecture is based on two basement levels: statistical and machine learning, that provide the necessary information to train the DL model which constitutes the third architectural level. The proposed DL method provides a modern approach to tackle fairness in RS. We can easily balance accuracy and fairness, or we can automatically select the optimum tradeoff. That is to say: the proposed method manages the inherent loss of accuracy when fairness is increased. Additionally, once the neural network is trained using demographic information, it can predict and recommend to users whose demographic information is unknown.

Results show adequate trends in the tested quality measures: improvement in fairness at the cost of an expected worsening in accuracy. The proposed machine learning-based heuristic approach and the DL model return similar quality results. Nevertheless, the proposed DL method does not need demographic information in the recommendation feed-forward process. It also can better balance and automatically balance fairness and accuracy.

The main contributions of the paper are:

- A novel Deep Learning based Collaborative Filtering algorithm that provides recommendations with an optimum balance between fairness and accuracy.
- Our proposed method does not require an initial knowledge of the users' demographic information.
- The proposed method relies on an original loss function and input data to balance fairness and accuracy. Also, it can manage the inherent loss of accuracy when fairness is increased, balancing accuracy and fairness of the recommendations.

Proposed future works are: a) architecture simplification, by removing the MF and transferring its functionality to the DL model, b) items and users minority indexes redefinition to better catch the minority versus non-minority differences, c) testing the methods behavior in a variety of CF datasets, d) extending the experiments to different demographic groups (nationality, profession, studies), and e) testing the architecture on not demographic groups (users that share minority preferences).

## Acknowledgment

This research was supported by the Ministerio de Ciencia e Innovación of Spain, grant number PID2019-106493RB-I00.

## References

[1] E. C¸ ano, M. Morisio, "Hybrid recommender systems: A systematic literature review," *Intelligent Data Analysis*, vol. 21, no. 6, 2017, pp. 1487–1524.

[2] A. Bellogín, P. Castells, I. Cantador, "Statistical biases in Information Retrieval metrics for recommender systems," *Information Retrieval Journal*, vol. 20, no. 6, 2017, pp. 606–634.

[3] R. Gao, C. Shah, "Toward creating a fairer ranking in search engine results," *Information Processing & Management*, vol. 57, no. 1, 2020, pp. 102138.

[4] M. Fatehkia, R. Kashyap, I. Weber, "Using Facebook ad data to track the global digital gender gap," *World Development*, vol. 107, 2018, pp. 189–209.

[5] N. S. Santos, A. Garc´ıa-Holgado, M. C. S´anchez-Go´mez, "Gender gap in the digital society: A qualitative analysis of the international conversation in the wyred project", in: *Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality*, TEEM'19, New York, NY, USA, 2019, pp. 518–524.

[6] I. Portugal, P. Alencar, D. Cowan, "The use of machine learning algorithms

in recommender systems: A systematic review," *Expert Systems with Applications*, vol. 97, 2018, pp. 205–227.

[7] M. Mendoza, N. Torres, "Evaluating content novelty in recommender systems," *Journal of Intelligent Information Systems*, vol. 54, no. 2, 2020, pp. 297–316.

[8] J. Bobadilla, A. Guti´errez, F. Ortega, B. Zhu, "Reliability quality measures for recommender systems," *Information Sciences*, vol. 442-443, 2018, pp. 145–157.

[9] M. Kunaver, T. Po˘zrl, "Diversity in recommender systems – A survey," *Knowledge-Based Systems*, vol. 123, 2017, pp. 154–162.

[10] M. de Gemmis, P. Lops, G. Semeraro, C. Musto, "An investigation on the serendipity problem in recommender systems," *Information Processing & Management*, vol. 51, no. 5, 2015, pp. 695–717.

[11] D. Kotkov, S. Wang, J. Veijalainen, "A survey of serendipity in recommender systems," *Knowledge-Based Systems*, vol. 111, 2016, pp. 180–192.

[12] K. Holstein, J. Wortman Vaughan, H. Daum´e, M. Dudik, H. Wallach, "Improving fairness in machine learning systems: What do industry practitioners need?," in: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, New York, NY, USA, 2019.

[13] R. Mehrotra, J. McInerney, H. Bouchard, M. Lalmas, F. Diaz, "Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems," in: P*roceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 2243–2251.

[14] J. L. Herlocker, J. A. Konstan, L. G. Terveen, J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems*, vol. 22, no. 1, 2004, pp. 5–53.

[15] A. Hernando, J. Bobadilla, F. Ortega, "A non negative matrix factorization for collaborative filtering recommender systems based on a Bayesian probabilistic model," *Knowledge-Based Systems*, vol. 97, 2016, pp. 188–202.

[16] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, "A survey on bias and fairness in machine learning", 2019. arXiv:1908.09635.

[17] R. Burke, N. Sonboli, A. Ordonez-Gauger, "Balanced neighborhoods for multi-sided fairness in recommendation," in: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, Vol. 81 of Proceedings of Machine Learning Research, PMLR, New York, NY, USA, 2018, pp. 202–214.

[18] J. Leonhardt, A. Anand, M. Khosla, "User fairness in recommender systems," in: *Companion Proceedings of the Web Conference 2018*, WWW '18, Republic and Canton of Geneva, CHE, 2018, pp. 101–102.

[19] M. D. Ekstrand, M. Tian, M. R. I. Kazi, H. Mehrpouyan, D. Kluver, Exploring author gender in book rating and recommendation, in: Proceedings of the 12th ACM Conference on Recommender Systems, RecSys '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 242–250.

[20] V. Tsintzou, E. Pitoura, P. Tsaparas, "Bias disparity in recommendation systems," arXiv:1811.01461.

[21] S. Yao, B. Huang, Beyond parity: Fairness objectives for collaborative filtering, *CoRR* abs/1705.08804.

[22] M. Mansoury, B. Mobasher, R. Burke, M. Pechenizkiy, "Bias Disparity in Collaborative Recommendation: Algorithmic Evaluation and Comparison," *ArXiv e-prints*.

[23] A. Chouldechova, A. Roth, "The frontiers of fairness in machine learning," *CoRR* abs/1810.08810.

[24] R. Mu, "A Survey of Recommender Systems Based on Deep Learning," *IEEE Access*, vol. 6, 2018, pp. 69009–69022.

[25] Z. Batmaz, A. Yurekli, A. Bilge, C. Kaleli, "A review on deep learning for recommender systems: challenges and remedies," *Artificial Intelligence Review*, vol. 52, no. 1, 2019, pp. 1–37.

[26] J. Bobadilla, F. Ortega, A. Guti´errez, S. Alonso, "Classification-based deep neural network architecture for collaborative filtering recommender systems," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 1, 2020, pp. 68–77.

[27] J. Bobadilla, S. Alonso, A. Hernando, "Deep Learning Architecture for Collaborative Filtering Recommender Systems," *Applied Sciences*, vol. 10, no. 7, 2020.

[28] J. Choo, S. Liu, "Visual Analytics for Explainable Deep Learning," *IEEE Computer Graphics and Applications*, vol. 38, no. 4, 2018, pp. 84–92.

[29] H. Wu, Z. Zhang, K. Yue, B. Zhang, J. He, L. Sun, "Dual-regularized matrix factorization with deep neural networks for recommender systems," *Knowledge-Based Systems*, vol. 145, 2018, pp. 46–58.

[30] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, T.-S. Chua, "Neural collaborative filtering," in: *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, Republic and Canton of Geneva, CHE, 2017, pp. 173–182.

[31] F. M. Harper, J. A. Konstan, "The movielens datasets: History and context," *ACM Transactions on Interactive and Intelligent Systems*, vol. 5, no. 4, 2015, pp. 1–19.

Jesús Bobadilla

Jesús Bobadilla received the B.S. and the Ph.D. degrees in computer science from the Universidad Politécnica de Madrid and the Universidad Carlos III. Currently, he is a lecturer with the Department of Applied Intelligent Systems, Universidad Politécnica de Madrid. He is a habitual author of programming languages books working with McGraw-"Hill, Ra-Ma and Alfa Omega publishers. His research interests include information retrieval, recommender systems and speech processing. He oversees the FilmAffinity.com research team working on the collaborative filtering kernel of the web site. He has been a researcher into the International Computer Science Institute at Berkeley University and into the Sheffield University. Head of the research group.

Raúl Lara-Cabrera

Raúl Lara-Cabrera received the M.Sc. and Ph.D. degrees in computer science from the University of Málaga, Spain, in 2013 and 2015, respectively. He is currently an Assistant professor with the Department of Sistemas Informáticos, Universidad Politécnica de Madrid, Spain. His main research interests include computational intelligence, machine learning, video games, and complex systems.

Ángel González-Prieto

Ángel González-Prieto received his Double B.S. in Computer Sciences and Mathematics from Universidad Autónoma de Madrid in 2014, his M.Sc. in Mathematics from the same university in 2015 and his Ph.D. in Mathematics from Universidad Complutense de Madrid in 2018. He has been postdoc at Instituto de Ciencias Matemáticas and, currently, he is Teaching Assistant at Universidad Politécnica de Madrid. His research interests include machine learning, deep learning, and algebraic geometry.

Fernando Ortega

Fernando Ortega was born in Madrid, Spain, in 1988. He received the B.S. degree in software engineering, the M.S. degree in artificial intelligence, and the Ph.D. degree in computer sciences from the Universidad Politécnica de Madrid, in 2010, 2011, and 2015, respectively, From 2008 to 2015, he was a Research Assistant with Intelligent Systems for Social learning and Virtual Environments Research Group. From 2015 to 2017, he was with BigTrueData Leading Machine Learning Projects. From 2017 to 2018, he was an Assistant Professor with the U-tad, Centro Universitario de Tecnología y Arte Digital. Since 2018, he has been an Assistant Professor with the Universidad Politécnica de Madrid. He is author of 30 research papers in most prestigious international journals. He leads several national projects to include machine learning algorithms into the society. His research interests include machine learning, data analysis, and artificial intelligence.

# NSL-BP: A Meta Classifier Model Based Prediction of Amazon Product Reviews

Pravin Kumar[1] , Mohit Dayal[2], Manju Khari[3]*, Giuseppe Fenza[4], Mariacristina Gallo[4]

[1] Indian Institute of Technology (ISM), Dhanbad (India)
[2] Ambedkar Institute of Advanced Communication Technology & Research (India)
[3] Netaji Subhas University of Technology, East Campus, Delhi (India)
[4] University of Salerno, Fisciano (SA), Italy

Received 11 April 2020 | Accepted 7 September 2020 | Published 6 October 2020

## Abstract

In machine learning, the product rating prediction based on the semantic analysis of the consumers' reviews is a relevant topic. Amazon is one of the most popular online retailers, with millions of customers purchasing and reviewing products. In the literature, many research projects work on the rating prediction of a given review. In this research project, we introduce a novel approach to enhance the accuracy of rating prediction by machine learning methods by processing the reviewed text. We trained our model by using many methods, so we propose a combined model to predict the ratings of products corresponding to a given review content. First, using k-means and LDA, we cluster the products and topics so that it will be easy to predict the ratings having the same kind of products and reviews together. We trained low, neutral, and high models based on clusters and topics of products. Then, by adopting a stacking ensemble model, we combine Naïve Bayes, Logistic Regression, and SVM to predict the ratings. We will combine these models into a two-level stack. We called this newly introduced model, NSL model, and compared the prediction performance with other methods at state of the art.

## Keywords

Combined Model, Logistic Regression, Machine Learning, Naïve Bayes, Stacking Model, SVM.

DOI: 10.9781/ijimai.2020.10.001

## I. Introduction

Nowadays, a large amount of customer reviews, available on every commercial site, provides valuable information about products but also impact the purchase decision of customers.

A recent survey of Ye, Q., Law [1] revealed that about 67.77% of customers are impacted by online reviews when they were making purchase decisions. However, rating prediction is also necessary because searching and comparing text reviews can be a headache for customers [2]. So users' reviews information should be merged, but a large number of reviews and unstructured text formats confuse users, making hard any decision. The star-rating, i.e., a star from 1 to 5 on any commercial site, can give a brief idea of product quality, more quickly than its text content. There are some interesting models that can predict user ratings from the text review [3]. Nevertheless, the rating prediction using reviews' text requires to face several challenges like human errors, vocabulary errors, and so on. The reviews may contain unreliable information increasing the quality of the task results. To get rid of these problems, we can rely on supervised machine learning techniques [4], such as text classification, which allows us to automatically classifying a document into a fixed set of classes according to its meaning. In this context, three different approaches for rating prediction could be applied: binary classification, multi-class classification, and logistic regression. The binary classification classifies a product as good or not, but using multiclass-classification and logistic regression, the customers are also informed about the level of quality of the product by giving a rating (for example, from 1 to 5).

This work proposes a new model (NSL) inspired by ensemble methods, which combine multiple existing models in order to obtain a better prediction result. In particular, adopted classifiers are Naive Bayes, Support Vector Machine (SVM), and Logistic Regression. The combination is made through a two-level stacking. All models have been trained by means of an Amazon dataset. In this sense, the approach also tries to face subsequent challenges:

1. The class imbalance: the dataset is relatively skewed in terms of class distribution.
2. In multi-class case, over-representation of 5-star ratings.

We overcome these issues by applying sampling techniques [6] to even out the class distribution. We dealt with the issue of class imbalance by investing in some balancing techniques [7].

Results show that the two most successful classifiers are Logistic regression and SVM. Still, Logistic regression gives better results than SVM. However, our combined model (the NSL model) gives the best results.

Machine learning algorithms are divided into supervised and unsupervised approaches. The first ones need the labeled data; the latter can be adopted with unlabeled data. Among supervised approaches, we will discuss about the text classification, which has been used in predicting the ratings. In particular, in Fig. 1, we describe the adopted process during text classification: from training data consisting of text documents we extract representing feature vectors

* Corresponding author.
E-mail address: manjukhari@yahoo.co.in

adopted during the training. Labeled training data helps the algorithm in discovering patterns between the input text and the respective rating. Finally, after the same preprocessing aiming to extract feature vectors, the constructed model is adopted on new data (i.e., test set) in order to discover new ratings.

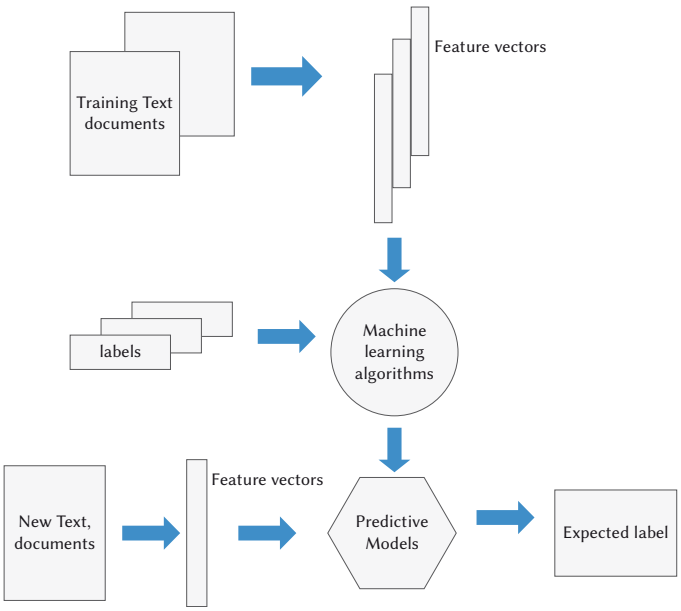



Fig. 1. Supervised Machine Learning.

## II. Theoretical Background

A classification task consists in the identification of a predefined class after the learning of a model through training data. The classification could be applied to different types of data. In the context of product ratings, we face a problem of text classification. Formally, the text classification tries to predict the best class c ε C for each document d ε D, where C is a fixed set of classes, and D is a collection of documents. Two types of classifications exist. Basing on the cardinality of C, we can distinguish between binary and multi-class classification (see Fig. 2 and Fig. 3). In particular, when there are only two classes, and each document belongs to one of the two classes, we face with binary classification (e.g. spam filtering in mails). In multi-class classification, there are more than two classes, and each document belongs to one of these classes. This is the case of rating reviews: classes go from 1-star to 5-star, where 1-star is considered the worst review class, and 5-star means the best review class.

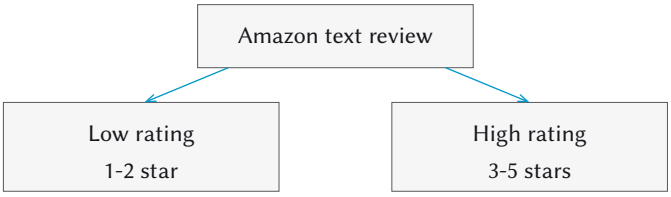



Fig. 2. Binary Classification.


Fig. 3. Multi-class Classification.

Sometimes, due to imperfection in datasets, other important steps (i.e., data cleaning, and resampling) must precede the model training process, as expressed in Fig. 4 and detailed as follows.


Fig 4. Text Classification Implementation.

Collected data should be cleaned in order to improve its quality: identify incomplete, incorrect, inaccurate and irrelevant parts of the data and then replacing, modifying, or deleting them.

The adopted dataset can be either balanced (Fig. 5 a) or imbalanced (Fig. 5 b). When classes are unequally distributed, the dataset is considered imbalanced.

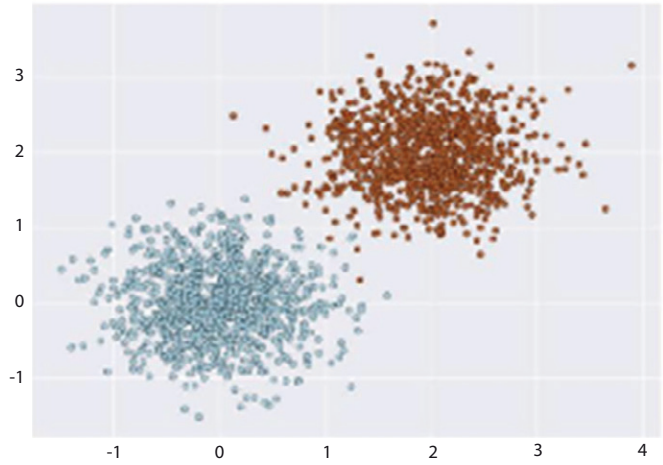



Fig. 5.a. Balanced Dataset

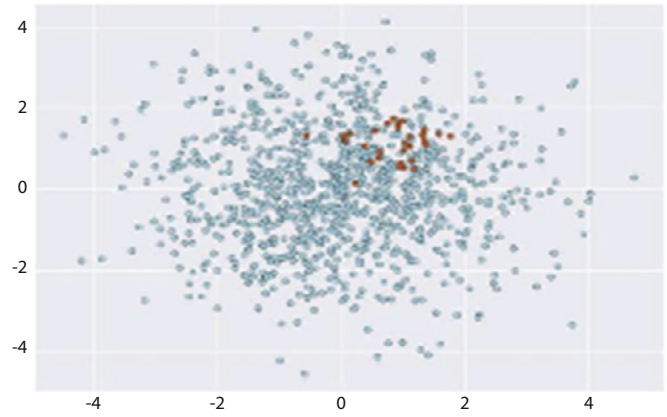



Fig. 5.b. Imbalanced Dataset.

To avoid imbalanced classification problems, various resampling [8], [9] methods can be applied. They aim at balancing data before its adoption. Imbalanced data can be treated by *under sampling* (i.e., reduction of items belonging to the most represented classes) or *oversampling* (i.e., addition of items for under-represented classes) processes.

### A. Text Classification Algorithms

Text classification can be made by classification algorithms (i.e., classifiers [23]). Here we present some of the most used classifiers.

#### 1. Naïve Bayes

Naïve Bayes classifiers belong to the family of probabilistic

classifiers that apply the Bayes's theorem [32], [33]. Specifically, the classifier calculates the probability by which the document belongs to a particular class. It is based on the MAXIMUM a Posteriori (MAP) estimator [24] that by means of the class prior probability assigns the best class to the document. The mathematical formula of the probability to predict a class c to a document d is defined in Eq. 1.

$$C_{map} = \arg\max_{c \in C} \hat{P}(c) \prod_{1<=k<=n_d} \hat{P}\left(t_k/_c\right) \quad (1)$$

Where, $\hat{P}(c)$ is the class prior probability, the probability that a document belongs to class c, $\hat{P}\left(t_k/_c\right)$ is the probability of a term t at position k in a document d from the class c, and $n_d$ is the number of terms in document d.

### 2. Support Vector Machine

Support Vector Machines (SVMs) are a class of supervised machine learning algorithms for binary classification problems. The key idea of SVM is to find the hyperplane $\Pi$ that separates the positive points from negative ones as wide as possible. Here W is normal to the plane. $W_1$ is normal to the plane $\Pi_1$ and $W_2$ is normal to the plane $\Pi_2$.

In Fig. 7, we have to reduce the margin of given hyperplanes so that we can be able to find the best hyperplane, which divides the data points accurately. Let us define the distance between the data point and the hyperplane as expressed in the following Equation.

$$y_i(w^T x_i + b)$$

If the distance is *less* than 1, the point is correctly classified; if the distance is *equal* to 1, the point is on the hyperplane; if the distance is greater than 1, the point is misclassified. On the basis of the distance, we can find out the best hyperplane to classify the data i.e depicted in Fig. 6, Fig. 7.

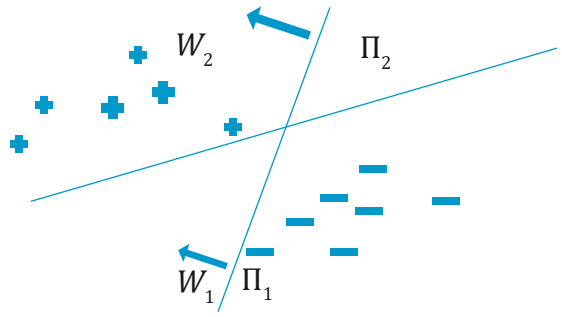



Fig. 6. Hyperplanes to divide the data.

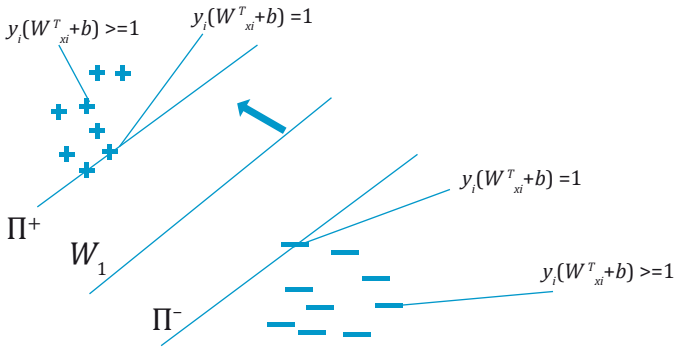



Fig. 7. Margin Hyperplane.

### 3. Logistic Regression

Logistic regression finds the plane that separates the different classes of data. There could be three interpretations of logistic regression, such as geometry, probability, and loss function, where all the optimization methods are related to each other, with some differences. In logistic regression, we are assuming that classes are linearly separable or almost linearly separable. If classes are not linearly separable, then we have to apply Feature Engineering on data. Feature Engineering consists of mathematical, trigonometry, or logarithmic functions [25].

### 4. Ensemble Methods

In addition to the described algorithms, there exist approaches that combine multiple base models in order to improve their overall performances. The combination of models can be realized through different aggregation criteria (i.e., bagging, boosting, stacking, and so on).

Our **proposed model is inspired to stacking (or stacked) approach**. It consists in a sequential method where all algorithms to combine are trained by the training set. Then, the new algorithm (the combined one, also considered as meta-classifier) is trained through the prediction outs of the other algorithms.

## III. Related Literature

In the area of customers' review classification, numerous solutions are available in the literature.

S. Wararat [10] classifies hotels' customer reviews written as open comments as positive or negative, using a binary classifier (i.e., opinion mining). This model, by adopting the Naïve Bayes technique, gives a prediction accuracy result of 94.37%. Lei et al. calculate each user's sentiment on products and take interpersonal sentimental influence and product reputation into consideration [11]. To make a correct rating prediction, authors fuse three factors into the recommender system [5]. Performance evaluation of the three sentimental factors is conducted on real-world data collected from Yelp. Baccouche et al. proposed a review data pre-processing and subsequent training of different classifiers (i.e., Multinomial Naïve Bayes, Bigram Multinomial Naïve Bayes, Trigram Multinomial Naïve Bayes, Bigram-Trigram Multinomial Naïve Bayes, Random Forest) [12]. In terms of accuracy, the Random Forest approach is the best one. Reddy et al. propose combined collaborative filtering of hierarchical topic models for integrating sentiment analysis [13]. By taking previous reviews, they predict future reviews of a given author. Kawamae introduces a simple supervised learning algorithm for semantic analysis for large text documents [14]. By using pointwise mutual information, the method involves issuing queries to a Web search engine.

Turney applied supervised machine learning classification algorithms to extract the semantic orientation of individual words extracted from a big corpus [15],[34].

To address the sentiment analysis for rating prediction, Kotsiantis et al. proposed graph-based semi-supervised learning algorithms [16]. The task is to give numerical ratings for unlabeled documents based on the perceived sentiment expressed by their text. In this paper, Goldberg et al. combine the LDA model and the association rules to extract the product features and corresponding words of reviews [17]. The authors used cross-validation to prune the extracted result. In this paper, to calculate how much important the word is for review, authors adopted an unsupervised approach and ranked the reviews.

Liu et al. propose a solution showing the meaning of phrases and sentences in vector space [18]. This approach is based on a vector construction through additive and multiplicative functions. Results show that multiplicative models are better than additive alternatives. Mitchell et al. use a vector space framework to represent sentences [19]. Tiroshi et al. propose a graph-based representation of the data in order to generate and self-populate features [20].

## IV. Proposed Methods

In terms of review rating prediction, we propose the NSL model inspired to ensemble method solutions that combine multiple classification models. In particular, we combine Naïve-Bayes, Logistic Regression, and SVM in a stacked way. The proposed model is shown in Fig. 8. It works as follows. Let X be the training data having n features. All three existing models are trained on X. The predictive output of each model is converted to a second level data, making each prediction a new feature for this second level. Then, we apply a meta-classifier training on this data. The meta-classifier result will be almost similar to the best of the three models.

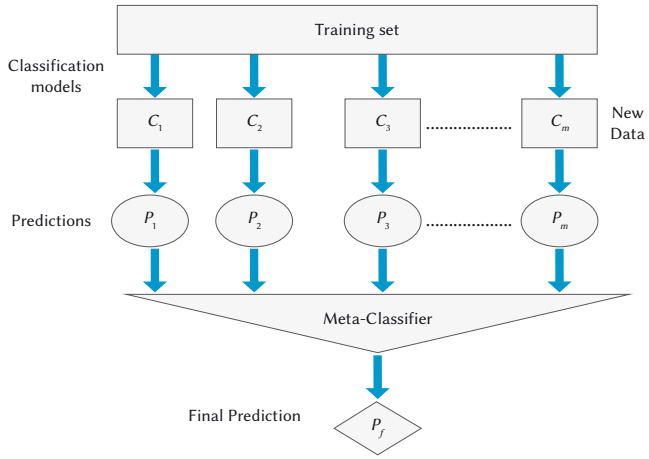



Fig. 8. NSL Model (Combined Model of Naive Bayes, Logistic Regression, SVM).

Inspired to Naïve-Bayes classifiers, we adopt the technique that predicts the best class for a document based on the probability that the terms in the document belong to the class. We took MAXIMUM a Posteriori (MAP) estimator described in Section 2.2.1.

From the Support Vector Machine, we took the minimization of the margin of hyperplane function as:

$$\text{Probabilistic } W^*, b^* = \arg\min_{w,b} \frac{\|W\|}{2} + c.\frac{1}{n}\sum_{i=1}^{n}\xi_i \tag{2}$$

$$m \arg in\ d = \frac{2}{\|W\|} \tag{3}$$

Where we have to maximize the margin and to draw the hyperplane that best divides the different data points of reviews.

$$W^*, b^* = \arg\min_{w,b} \frac{2}{\|W\|} \tag{4}$$

Such that $y_i(w^T x_i^{+b}) >= 1$ for all $x_i$. It is the hinge loss of SVM. Such that $y_i(w^T x_i^{+b}) > 1 - \xi$, where $\xi >= 0$

From the Logistic regression, we took the geometrical part; in this function, we minimize the distance of every point from the hyperplane and search for the hyperplane, which gives the best results:

$$W * = \arg\min_{W} \sum_{i=1}^{n} -y_i \log p_i - (1 - y_i)\log(1 - p_i) \tag{5}$$

Where, $y_i$ is +1 or 0: positive and negative points, respectively. Since $p_i = \sigma(w^T x_i)$ is a sigmoidal function, we have to minimize the probabilistic distance function. Then, the result of the meta-classifier will be the final prediction of that data point on the majority voting basis.

In order to preserve the figures' integrity across multiple computer platforms, we accept files in the following formats: .EPS/.PDF/.PS/.AI. All fonts must be embedded or text converted to outlines in order to achieve the best-quality results.

### A. Text Preprocessing

Data Cleaning and Data Resampling are two important methods of Text Preprocessing.

The objective of data cleaning consists of: (i) punctuation discarding, (ii) number discarding, (iii) lower-casing of the text (iv) extra whitespace removing, and (v) stop word (like "is", "are", "a", "and", etc.) removing. It simplifies data and makes classification more accurate.

It is observed that the review data has many duplicate entries, so remove duplicates can unbias results in data analysis. Among available methods to remove duplicates, our choice consists of removing multiple reviews of the same user at the same time.

Before starting subsequent steps, we plotted data to recognize resampling needs and adopted suitable solutions in terms of data balancing.

In terms of representation of data, our solution treats the training corpus as a Bag of Words [21] and turned it in numerical feature vectors [22] using the *CountVectorizer* method, described following. So, given text reviews, the objective consists of extracting vectors of d dimension and finding a plane that represents them. Let be:

i) $r_1, r_2, r_3 \ldots\ldots\ldots r_n$, the reviews,

ii) $v_1, v_2, v_3 \ldots\ldots\ldots v_i$, the vectors of reviews in d dimension space.

iii) If Similarity $(r_1, r_2)$>Similarity $(r_1, r_3)$ then distance $(v_1, v_2)$ < distance $(v_1, v_3)$.

iv) If $r_1$ and $r_2$ are more similar then $v_1$, $v_2$ are more closed.

Regarding the value for each feature (i.e., a word), since using only its occurrence could be poor, the TF-IDF [29] Transformer method has been used in order to obtain its TF-IDF value. Let be:

$TF(W_i, r_j)$ = Number of times $w_i$ occur in $r_j$/total number of words in $r_j$.

$$IDF(w_i, D_C) = \log\left({}^{N}/_{n_i}\right)$$

where $N$ is the total number of documents, and $n_i$ the number of documents which contain $w_i$.

$$n_i <= N => {}^{N}/_{n_i} >= 1 \Rightarrow \log\left({}^{N}/_{n_i}\right) >= 0$$

If $w_i$ is much frequent in the corpus, then the IDF will be very low. If $w_i$ is a rare word, then IDF will be high.

### B. Training and Classification

The classifier, during the training phase, learns the mapping between a document and a class. Subsequently, it is able to classify new documents accurately.

A Latent Dirichlet Allocation (LDA) [27] task and a k-means clustering [28] step precede the training process. The LDA divides all reviews based on topic discussed within the text. We span amazon reviews from 1-star to 5-star and we bucket reviews by following criteria:

- Low: 1-2 star (if low > 80%, then 1 star, otherwise 2 stars)
- Neutral: 3 star (if neutral ≅ 50)
- High: 4-5 star (if high ≥ 80%, then 5 star, if high < 80% and high > 60%, then 4 stars)

K-means, based on the TFIDF matrix, groups documents into N clusters. Within each cluster, we count top occurring terms. By using LDA and TFIDF matrix, we attempt to extract N topics from our collection of documents. K-means forces each review to belong to only one cluster, but LDA allows a review to have many topics associated with it.

The algorithm for training of the model and the subsequent classification of new data is resumed in Algorithm 1.

**Algorithm 1**. Construction and Adoption of Classification Model

**Inputs** Training data D= $\{x_i, y_i\}$ $i$=1 to m $\{x_i^{\varepsilon} R^n, y_i \varepsilon \gamma\}$

**Output** *An ensemble classifier H*

1. Step1:Training of data
2. Step2: Using TF-IDF matrix do K-MEANS clustering on the review documents with 9 clusters
3. Step3: Using TF matrix use LDA for Extract top topics from clusters
4. Step4: learn first level classifiers
5. For t ← 1 to T do
6. Learn a base classifier h based on D
7. End for
8. Step5: construct new data set from D
9. For $i$←1 to m do
10. Construct a new data set that contains

$\{x_i, y_i\}$, where $x_i = \{h_1(x_1), h_2(x_2), \ldots\ldots\ldots\ldots h_i(x_i)\}$

11. End for
12. Step6: Learn a second level classifier
13. Learn a new meta-classifier h' based on newly constructed data set with functions
14. Training of meta-classifier to classify the data

Step 7: From the Naïve-Bayes MAXIMUM a Posteriori Function is used

$$Cmap = \arg\max_{c \epsilon C} \hat{P}(c) \prod_{l<=k<=n_d} \hat{P}\left(tk/_c\right)$$

Step 8: From the support vector machine Minimization of margin of hyperplane is used

Function for Minimization of margin of hyperplane

$$w^*, b^* = \arg\min_{w,b} \frac{\|W\|}{2} + c.\frac{1}{n}\sum_{i=1}^{n} \xi_i$$

Step 9: From the Logistic regression probabilistic part, we minimize the distance of every point from the hyperplane, and search for the hyperplane, which gives the best results. Function for Minimization of distance from hyperplane

$$w^* = \arg\min_{W} \sum_{i=1}^{n} -y_i \log p_i - (1-y_i)\log(1-p_i)$$

15. Step 10: predict the class based on the majority of voting
16. Return

## V. Experiment and Result Analysis

The NSL model, generated as described in the previous sections, is evaluated through experimentation on a real dataset as expressed following.

### A. Dataset

Experimental analysis has been done on the freely available Amazon dataset "Home and Kitchen"[1]. The dataset contains 500k kitchen product reviews from May 1996 to July 2014. Each review contains product id, user id, profile name, helpfulness rating (e.g., 2/3), time, summary, text.

[1] http://snap.stanford.edu/data/amazon/productGraph/categoryFiles/reviews

### B. Evaluation

Regarding the validation of the model, we apply the k-fold method. In k-fold validation [26], we divide the training data into many subsets. By dividing our training data into N sets, we hold the Nth set for validation. We have three models, and, as already defined, we obtain the prediction from each one. Obtained predictions are collected in the out-of-sample prediction matrix. This matrix is used as a second level training data to obtain the final prediction. Second level training will select the best of the first level prediction models. By using out-of-sample prediction, we still have large data to train the second-level model. In the meta-classifier, we use various loss functions based on each adopted model, as explained in Section 4.2.

### C. Experimental Results

After the creation of the TFIDF matrix, we apply K-means clustering to extract the clusters, as expressed in Fig. 9. The distribution of the topics extracted through the LDA application is shown in Fig. 10. Table I lists the first 10 topics and 15 words per topic associated with them.

The resulting TF-IDF matrix has thousands of attributes, so it is very challenging to show them graphically since we can plot up to 3-dimensional only. We use the Latent Semantic Analysis [30] based on the singular value decomposition [32] to reduce the dimensionality of the matrix. We then use T-SNE [31] to represent our data as best as possible in 2-dimensions.

In Fig. 11, we extract the top 24 words in the "Low" category, 12 of them with red color are not associated, while 12 with green color are associated (i.e., terms more correlated with this type of category). Fig. 12 and Fig. 13 do the same for "Neutral" and "High" categories, respectively.

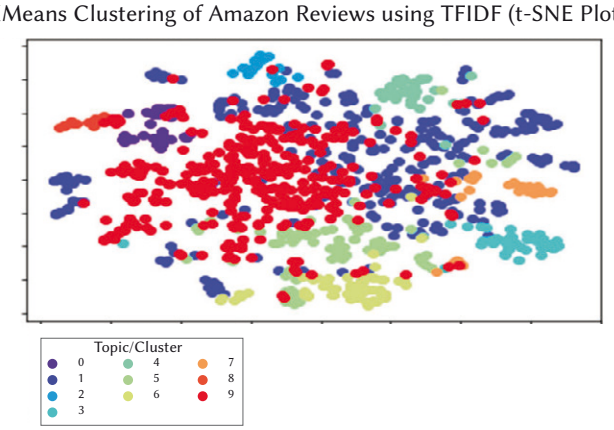



Fig. 9. K-means clustering of Amazon Reviews using TFIDF.

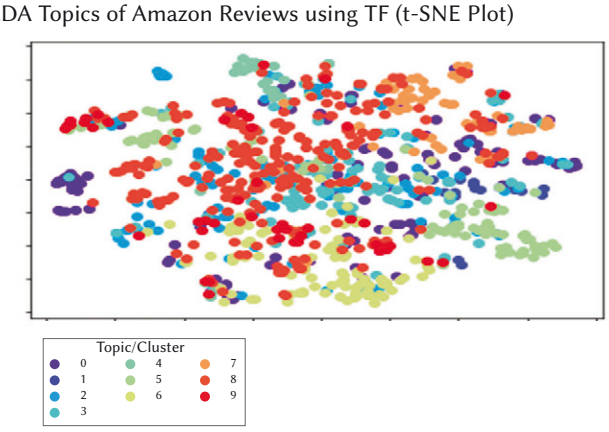



Fig. 10. LDA Topics of Amazon using TF.

TABLE I. Extracted Topics

| Topic | Associated words |
|---|---|
| Topic #0 | use, time, make, machin, work, pot, clean, just, tri, like, unit, onli, juic, turn, blender |
| Topic #1 | unit, fan, air, set, room, high, heat, assembl, low, veri, nois, heater, loud, turn, control |
| Topic #2 | just, work, use, bag, thing, review, like, time, realli, don't, i'm, veri, say, product, read |
| Topic #3 | product, year, replac, amazon, purchas, use, return, new, buy, time, review, just, month, work, did |
| Topic #4 | cut, blade, knife, grinder, grind, use, knive, sharp, edg, steel,handl, slice, veri, good, hand |
| Topic #5 | vacuum, clean, use, floor, bed, like, veri, carpet, mattress, brush, doe, power, attach, cord, cleaner |
| Topic #6 | coffee, cup, water, filter, use, make, glass, hot, tea, brew, maker, drink, like, pour, mug |
| Topic #7 | pan, cook, use, oven, stick, heat, egg, bread, toaster, oil, food, clean, toast, grill, set |
| Topic #8 | use, like, look, veri, just, nice, good, wash, realli, don't, fit, size, hold, make, color |
| Topic #9 | lid, water, use, pillow, open, plastic, size, small, like, fit, bowl, contain, kettl, jar, food |

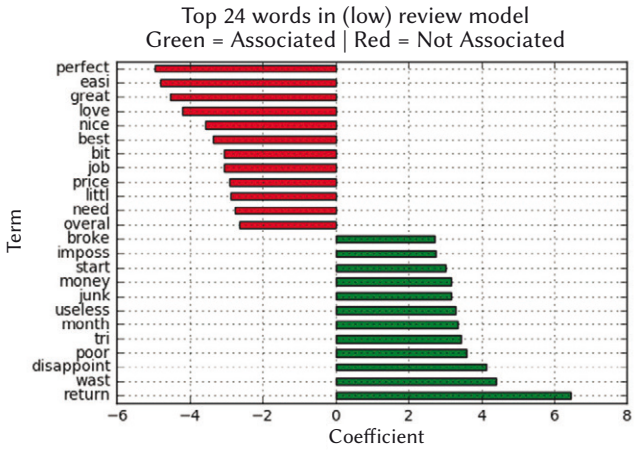



Fig. 11. Top 24 words in (low) review model.

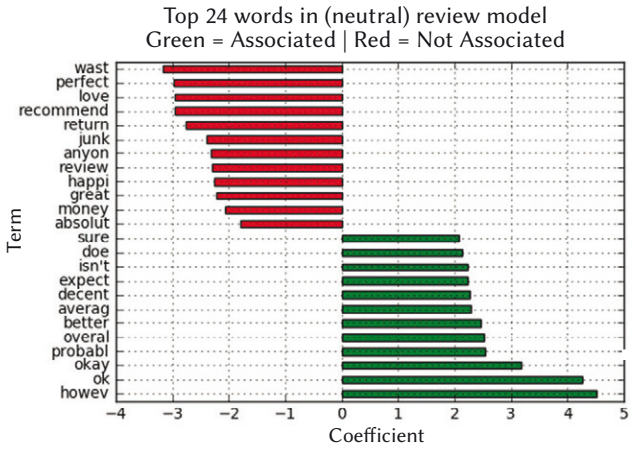



Fig. 12. Top 24 words in (neutral) review model.

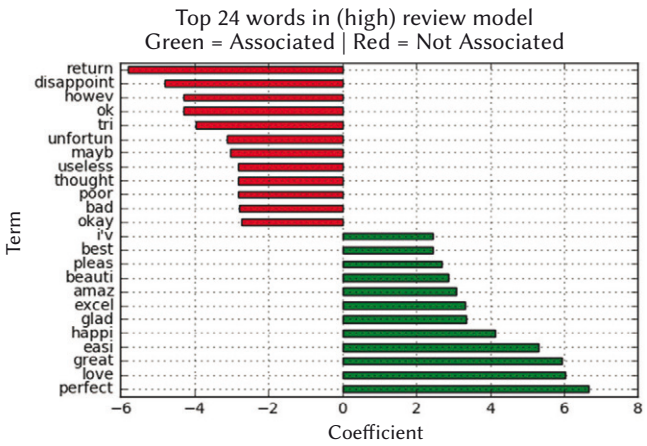



Fig. 13. Top 24 words in (high) review model.

Applying the learning model in new data rows (i.e., data does not used during the training), four categories of results can be encountered:

1. **True Positive (TP)**: prediction accurately predicted as positive
2. **True Negative (TN)**: prediction accurately predicted as negative
3. **False Positive (FP)**: prediction incorrectly predicted as positive
4. **False Negative (FN)**: prediction incorrectly predicted as negative

The prediction Accuracy of the classifier can be calculated as

$$\text{Error Rate} = \frac{FP+FN}{TP+TN+FN+FP}$$

$$\text{Accuracy} = 1-\text{Error Rate} = \frac{TP+TN}{TP+TN+FN+FP}$$

The error rate simply calculates the ratio between the number of wrong predictions made by our classifier and total number of test cases. From Table II to Table V, accuracy is reported for every adopted classifier. Our NSL model gives the highest accuracy with respect to other models. From Fig. 14, Fig. 15, Fig. 16 and Fig. 17, we compare accuracy of all the classifiers with our NSL model.

TABLE II. Accuracy Prediction of LR Model

| LR Model | Accuracy |
|---|---|
| **Low rating** | 75.8 % |
| **Neutral rating** | 68.6% |
| **High rating** | 77.8% |

TABLE III. Accuracy Prediction of SVM Model

| SVM Model | Accuracy |
|---|---|
| **Low rating** | 74.6 % |
| **Neutral rating** | 65.8% |
| **High rating** | 77.4% |

TABLE IV. Accuracy Prediction of NB Model

| NB Model | Accuracy |
|---|---|
| **Low rating** | 70.1 % |
| **Neutral rating** | 66.8% |
| **High rating** | 69.1% |

TABLE V. Accuracy Prediction of NSL (Combined Model)

| NSL Model | Accuracy |
|---|---|
| **Low rating** | 76.2 % |
| **Neutral rating** | 69.2% |
| **High rating** | 78.5% |

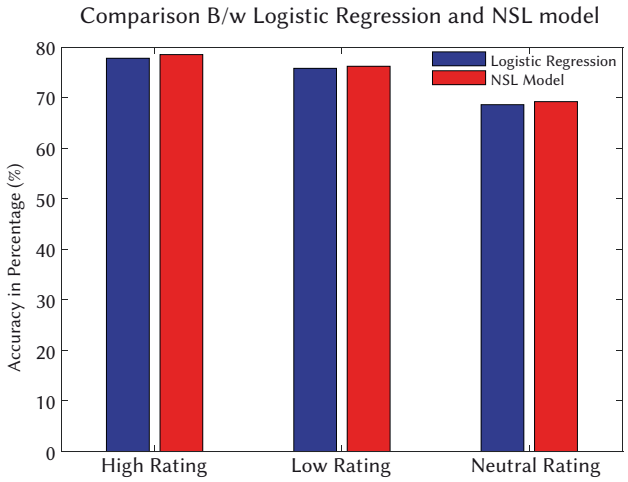



Fig. 14. Combined NSL model and Logistic Regression.

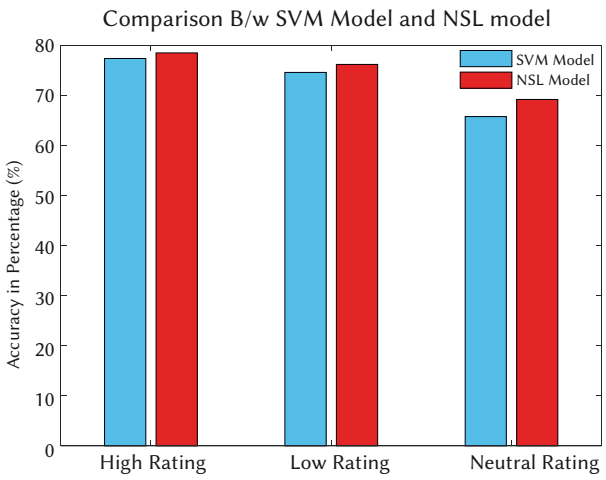



Fig. 15. Combined NSL Model and SVM.

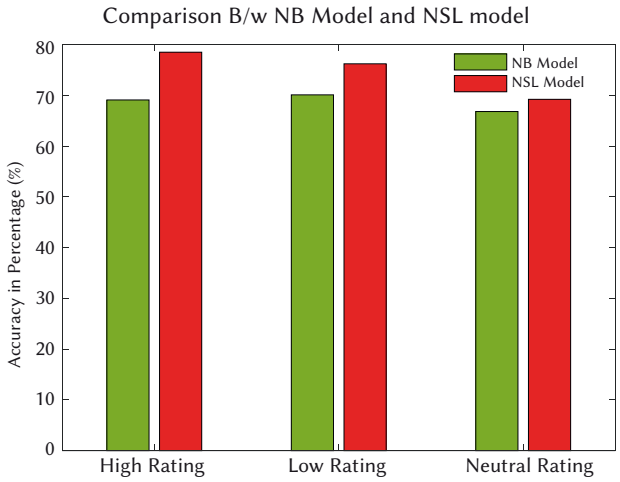



Fig. 16. Combined NSL Model and Naïve Bayes.

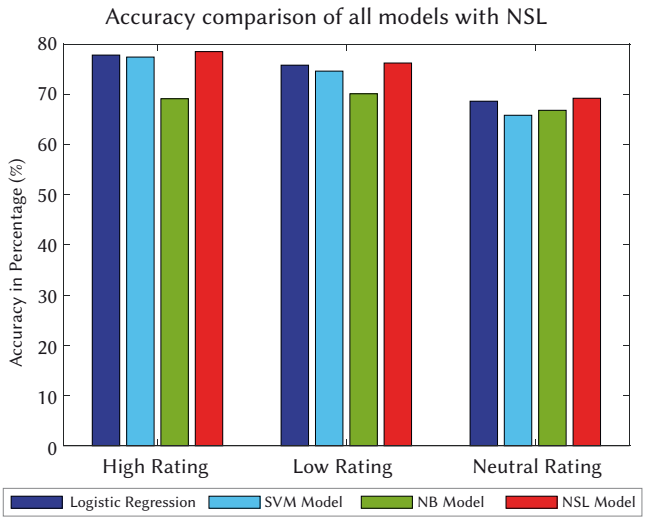



Fig. 17. Accuracy Comparison of all the classifiers.

## VI. Conclusion And Future Scope

In this work, we present NSL, a new classification model as a 2-level combination of existing ones (namely, Naïve Bayes, SVM, and logistic Regression classifier). For data preprocessing the underlying method Latent Dirichlet Allocation (LDA and k-means clustering algorithm are used. The LDA divides all reviews based on topic discussed within the text. All the used existing models are combined in stack manner. The predictive output of each model is converted to a second level data, making each prediction a new feature for this second level. After applying the meta classifier on training data, it is observed that the outcome of meta classifier is almost similar to the best underlying model. NSL successfully predicts a user's numerical rating from its review text content. The performance of the classifier is based on the necessity in terms of effectiveness, and it is also concerned with the number of features to be taken care during training and validation phase.

Experimentation has been done by means of "Home and Kitchen" dataset from Amazon. After a preprocessing aiming to balance the dataset, we train and test all four models and compare their performances. The outcomes of this experimental work are based on one type and category of the dataset. Further this classification can be used with other category of the dataset like food product, electronics appliances, delivery rating of product given by users. According to the evaluation metric, SVM gives the best result among multiclass classifiers but Logistic regression gives somehow better result than SVM. However, our proposed model overcomes, in terms of accuracy, all other models. Further scope of the underlying classifier extends for web page classification, electronic-mail classification, detection and classification of unauthorized signatures by combining of Hidden Naive Bayes and NBTree to decrease the Error rate of the classifier. When these enhancements are incorporated in the underlying classification system, it would help further improve the performance and be useful for applications meant for the explicit classification system.

## References

[1] Y. Qiang, R. Law, B. Gu, and W. Chen. "The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings." *Computers in Human behavior* 27, no. 2, pp. 634-639, 2011.

[2] G. Gayatree, N. Elhadad, and A. Marian. "Beyond the stars: improving rating predictions using review text content." *In WebDB*, vol. 9, pp. 1-6. 2009.

[3] B. Stefano, A. Esuli, and F. Sebastiani. "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining." *In Lrec*, vol. 10, no. 2010, pp. 2200-2204. 2010.

[4] K. B. Sotiris, I. Zaharakis, and P. Pintelas. "Supervised machine learning:

A review of classification techniques." *Emerging artificial intelligence applications in computer engineering*. Vol. 160, no. 1, pp. 3-14, 2007.

[5] L. Pasquale, M. D. Gemmis, and G. Semeraro. "Content-based recommender systems: State of the art and trends." *In Recommender systems handbook,* pp. 73-105. Springer, Boston, MA, 2011.

[6] C. G. William, 2007. "Sampling techniques". *John Wiley & Sons*, 2007.

[7] T. F. Brian, J. H. Patterson, and W. V. Gehrlein. "A comparative evaluation of heuristic line balancing techniques." *Management science* 32, no. 4 (1986): 430-454.

[8] Y. P. Chaubey, "Resampling-based multiple testing: Examples and methods for p-value adjustment." (1993): 450-451.

[9] D. M. Hawkins, 2004. The problem of overfitting. *Journal of chemical information and computer sciences*, 2004, 44(1), pp.1-12.

[10] S. Wararat. "The analysis and prediction of customer review rating using opinion mining." *In 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA)*, pp. 71-77. IEEE, 2017.

[11] L. Xiaojiang, X. Qian, and G. Zhao. "Rating prediction based on social sentiment from textual reviews." *IEEE transactions on multimedia* 18, no. 9 (2016): 1910-1921.

[12] B. Moez, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. "Sequential deep learning for human action recognition." *In International workshop on human behavior understanding*, pp. 29-39. Springer, Berlin, Heidelberg, 2011.

[13] S. C. Reddy, K. U. Kumar, J. D. Keshav, B. R. Prasad, and S. Agarwal. "Prediction of star ratings from online reviews." *In TENCON 2017-2017 IEEE Region 10 Conference*, pp. 1857-1861. IEEE, 2017.

[14] K. Noriaki. "Predicting future reviews: sentiment analysis models for collaborative filtering." *In Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 605-614. 2011.

[15] H. Jiawei, and KC-C. Chang. "Data mining for web intelligence." *Computer* 35, no. 11 (2002): 64-70.

[16] P. D. Turney, and M. L. Littman. "Unsupervised learning of semantic orientation from a hundred-billion-word corpus." *arXiv preprint* cs/0212012 (2002).

[17] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas. "Supervised machine learning: A review of classification techniques." *Emerging artificial intelligence applications in computer engineering* 160, no. 1 (2007): 3-24.

[18] A. B. Goldberg, and X. Zhu. "Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization." *In Proceedings of TextGraphs: The first workshop on graph based methods for natural language processing*, pp. 45-52. 2006.

[19] L. LiZhen, W. Wang, and H. Wang. "Summarizing customer reviews based on product features." *In 2012 5th International Congress on Image and Signal Processing*, pp. 1615-1619. IEEE, 2012.

[20] M. Jeff, and M. Lapata. "Vector-based models of semantic composition." *In proceedings of ACL-08: HLT,* pp. 236-244. 2008.

[21] T. Amit, S. Berkovsky, M. A. Kaafar, D. Vallet, T. Chen, and T. Kuflik. "Improving business rating predictions using graph based features." *In Proceedings of the 19th international conference on Intelligent User Interfaces*, pp. 17-26. 2014.

[22] Y. Zhang, R. Jin, and Z.H. Zhou. "Understanding bag-of-words model: a statistical framework." *International Journal of Machine Learning and Cybernetics* 1, no. 1-4 (2010): 43-52.

[23] A. Tiroshi, S. Berkovsky, M. A. Kaafar, D. Vallet, T. Chen, and T. Kuflik. "Improving business rating predictions using graph based features." *In Proceedings of the 19th international conference on Intelligent User Interfaces*, pp. 17-26. 2014.

[24] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas. "Supervised machine learning: A review of classification techniques." *Emerging artificial intelligence applications in computer engineering* 160, no. 1 (2007): 3-24.

[25] D. M. Greig, B. T. Porteous, and A. H. Seheult. "Exact maximum a posteriori estimation for binary images." *Journal of the Royal Statistical Society: Series B (Methodological)* 51, no. 2 (1989): 271-279.

[26] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection." *In Ijcai*, vol. 14, no. 2, pp. 1137-1145. 1995.

[27] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao. "Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey." *Multimedia Tools and Applications* 78, no. 11 (2019): 15169-15211.

[28] A. K. Jain, "Data clustering: 50 years beyond K-means." *Pattern recognition letters* 31, no. 8 (2010): 651-666.

[29] J. Ramos, "Using tf-idf to determine word relevance in document queries." *In Proceedings of the first instructional conference on machine learning*, vol. 242, pp. 133-142. 2003.

[30] T. K. Landauer, P. W. Foltz, and D. Laham. "An introduction to latent semantic analysis." *Discourse processes* 25, no. 2-3 (1998): 259-284.

[31] E. R. Henry, and J. Hofrichter. "Singular value decomposition: Application to analysis of experimental data." *In Methods in enzymology*, vol. 210, pp. 129-192. Academic Press, 1992.

[32] L. V. Maaten, and G. Hinton. "Visualizing data using t-SNE." *Journal of machine learning research* 9, no. Nov (2008): 2579-2605.

[33] L. E. Sucar, "Probabilistic graphical models." *Advances in Computer Vision and Pattern Recognition. London: Springer London.* doi 10 (2015): 978-1.

[34] J. Friedman, T. Hastie, and R. Tibshirani. The elements of statistical learning. Vol. 1, no. 10. *New York: Springer series in statistics*, 2001.

Pravin Kumar

Pravin Kumar is Member of Technical Staff Engineer at Mavenir System, Bangalore. He is currently working in Research and Development Department as a Research and Product development Engineer. Prior to Mavenir he Worked as a senior Research Engineer includes Autonomous Driving, Image processing and Algorithm Design and Analysis for Automotive components to make Driving Intelligent and Safe. He received his master's degree in Computer Science and Engineering from Indian Institute of Technology (ISM) Dhanbad. He holds a bachelor's degree in Computer Science & Engineering from University Of Pune. His research interests include machine learning, Internet of Things, 4G/5G Protocol Stack Development, Algorithm Design And Analysis and Swarm intelligence.

Mohit Dayal

Mohit Dayal is Technical committee Member of IEEE INDIaCom international conference, Delhi and Editorial member of International journal of Recent Advances in Science and Technology. He is currently working in Central University of Haryana, Mahendergarh, Haryana as an assistant professor in computer science and engineering Department. He received his master's degree in Information Security from Ambedkar Institute of Advanced Communication Technologies & Research of Guru Gobind Singh Indraprastha University, Delhi. He holds a bachelor's degree in Computer Science & Engineering from Guru Gobind Singh Indraprastha University, Delhi. His research interests include machine learning, Internet of Things, Big Data, Web Application attacks and information security.

Manju Khari

Manju Khari an Assistant Professor in Netaji Subhas University of Technology, East Campus, Delhi, India formerly Ambedkar Institute of Advanced Communication Technology and Research, Under Govt. Of NCT Delhi affiliated with Guru Gobind Singh Indraprastha University, Delhi, India. She is also the Professor- In-charge of the IT Services of the Institute and has experience of more than twelve years in Network Planning & Management. She holds a Ph.D. in Computer Science & Engineering from National Institute Of Technology Patna and She received her master's degree in Information Security from Ambedkar Institute of Advanced Communication Technology and Research, formally this institute is known as Ambedkar Institute Of Technology affiliated with Guru Gobind Singh Indraprastha University, Delhi, India. Her research interests are software testing, software quality, software metrics, information security, optimization and nature-inspired algorithm. She has 70 published papers in refereed National/International Journals & Conferences (viz. IEEE, ACM, Springer, Inderscience, and Elsevier), 06 book chapters in a springer. She is also co-author of two books published by NCERT of Secondary and senior Secondary School.

Giuseppe Fenza

Giuseppe Fenza is currently an Associate Professor in Computer Science at Department of Management and Innovation Systems, University of Salerno, Italy. He received the Ph.D. degree in Computer Sciences at the University of Salerno, Italy, in 2009. From 2009, his research interests ware mainly focused on Knowledge Extraction from unstructured resources defining intelligent systems based on the combination of techniques from Soft Computing, Semantic Web, areas in which he has many publications. He was deeply involved in several EU and Italian Research and Development projects focused on Situation Awareness, Service Discovery, Enterprise Information Management and e-Commerce. He serves as Associate Editor in international journals, such as: Neurocomputing, International Journal of Grid and Utility Computing, International Journal of Engineering Business Management. He has published extensively about: Fuzzy Decision Making, Ontology Elicitation, Situation and Context Awareness, Semantic Information Retrieval. Recently, he is working in the field of Big Data, Social Media Analytics, and Web Intelligence by proposing novel methods for instance to support microblog summarization, time-aware information retrieval and recommendations extraction. In 2017, he co-founded of Riatlas srl, a company that operates in the field of e-health. From 2018, he is participating in research projects focused in the area of big data analysis and analytics applied to Industry 4.0 and Cyber Physical Systems (CPS), that are: Leonardo 4.0 funded by the national Ministry of Education, Universities and Research; and CPS4EU funded by European ECSEL-IA H2020.

Maria Cristina Gallo

Maria Cristina Gallo received her Master Degree in Computer Science at the University of Salerno, Italy, in 2009. From 2009 to 2017, she collaborates to several research initiatives and projects mainly focused on Computational Intelligence, Data Mining, Ontology Learning e Semantic Information Retrieval in different domain, such as health, e-commerce, and enterprise. Recently, she has worked in the field of social Media Analytics and Semantic Web to study users' interests, characteristics of their posts, and potential cyclic nature of both of them. From 2017 until now, she is a PhD student in Big Data Management at the University of Salerno and she is involved in a project regarding Pattern Recognition and anomaly detection in data streams.

# Dynamic Generation of Investment Recommendations Using Grammatical Evolution

Carlos Martín, David Quintana*, Pedro Isasi

Department of Computer Science and Engineering, Universidad Carlos III de Madrid, Leganés (Spain)

Received 6 July 2020 | Accepted 20 March 2021 | Published 22 April 2021

## Abstract

The attainment of trading rules using Grammatical Evolution traditionally follows a static approach. A single rule is obtained and then used to generate investment recommendations over time. The main disadvantage of this approach is that it does not consider the need to adapt to the structural changes that are often associated with financial time series. We improve the canonical approach introducing an alternative that involves a dynamic selection mechanism that switches between an active rule and a candidate one optimized for the most recent market data available. The proposed solution seeks the flexibility required by structural changes while limiting the transaction costs commonly associated with constant model updates. The performance of the algorithm is compared with four alternatives: the standard static approach; a sliding window-based generation of trading rules that are used for a single time period, and two ensemble-based strategies. The experimental results, based on market data, show that the suggested approach beats the rest.

## Keywords

Dynamic Strategy, Evolutionary Computation, Finance, Grammatical Evolution, Structural Change, Trading.

DOI: 10.9781/ijimai.2021.04.007

## I. Introduction

FINANCIAL markets are subject to structural changes which cause investment rules that were profitable in the past to lose their effectiveness over time. This characteristic of market dynamics requires the implementation of mechanisms that detect structural changes and update investment strategies accordingly.

From Allen and Karjalainen [1] influential work on evolution of trading rules using Genetic Programming (GP) [2], many authors have followed with contributions based on the same technique or on the use of Grammatical Evolution (GE) [3]. Among them, [4]-[11].

The main disadvantage of the rules generated by the standard algorithm is the fact that they are static and, therefore, do not take into account the prevalent structural changes. Given the situation, the range of possibilities would have two extremes: either holding the same model through time or updating it with any new piece of information to keep up with the evolution of the price generation process.

Even though one might think that the constant update of the model is likely to be more appropriate than the static alternative, in practice it is often the case that the latter results in an increase in transaction costs that erodes completely the additional return obtained from better market timing.

This study improves the standard approach presenting an alternative that commutes between an active model and a candidate one. The algorithm has a hysteresis component that limits overtrading, while maintaining the ability to change the active model to cope with changes in the market price generation mechanism.

* Corresponding author.

E-mail address: dquintan@inf.uc3m.es

This research contributes to the state of the art improving the traditional use of Grammatical Evolution in this context. GE and GP-based approaches are subject to well-known limitations [12], such as the difficulties to outperform the market in the face of strong upward trends, but are popular due the advantages that they offer, such as flexible representation. While there are many potential trading algorithms based on a wide variety of techniques [13]-[15], benchmarking them would require a separate study that goes beyond the aim of this one.

The rest of the document is structured as follows: first, we provide a brief overview of the relevant literature on GE for algorithmic trading. That will be followed by a description of the canonical GE-based static approach and then the introduction of the proposed solution in section IV. The experimental analysis used to evaluate the approach will be reported in section V. Finally, we will devote the last section to summary and conclusions.

## II. Literature on GE for Trading

As we already mentioned, the efforts to obtain profitable trading rules using technical indicators with flexible representation is not new. The seminal contribution by Allen and Karjalainen [1] using GP paved the way for a substantial amount of related works. These extend the mentioned study suggesting new sets of functions and terminals, evaluation methods or investing universes.

In addition to these variations, other authors have explored the possibility of relying on a different core algorithm, GE. This approach, also under the framework of Evolutionary Computation, shares the main advantages of GP in this context and has been widely used in finance and economics [16]. According to a recent study [17], this technique seems to be more robust and to generate trading rules that are simpler and, therefore, easier to interpret. Hence, our decision to rely on GE as the core algorithm.

Among the earliest studies focused on GE for trading purposes we can highlight the pioneering one authored by Brabazon and O'Neill [18]. This work explores the possibility of generating investment rules for the money market using GE. To that end, these researchers relied on a limited set of indicators together with a risk management mechanism that offered encouraging results. The same year, Dempsey et al. [19] also used GE to evolve investment rules based on technical indicators for the Nikkei 225 and S&P 500 indices. This time, the authors highlight the importance of limiting transaction costs and present a strategy to reduce the number of trading signals by means of a decay constant that controls the investment size of the trades. Their results suggest that the approach is useful. The Nikkei 225 index was beat by a wide margin while the American benchmark seemed to be significantly more difficult to exploit.

Contreras et al. [20] tested the potential of GE to produce profitable trading rules in the Spanish stock market using technical indicators. These researchers compare the profitability of the resulting strategies vs a GA-based alternative that they introduced in a previous work [21] and report profits of 14% vs losses of 20% by the GA-based strategy. These same authors presented more recently two extended versions. The first one, a hybrid one that follows a multi-objective approach [22], includes multi-strategies to limit unforeseen losses and offered good performance in their experimental analysis. The second one [23], was used to explore the feasibility of using meta-GE approximation. The approach, which relies on two overlapped instances of grammatical evolution, uses a combination of macroeconomic, fundamental and technical indicators to generate trading rules. Once again, the system, which promoted more robust and lower-risk portfolios, resulted in promising results.

Schmidbauer et al. [24] introduce a GE-based trading rule selection framework that considers robustness. To that end, they developed a multi-objective fitness test that considers both observed series and synthetic ones generated using bootstrap. They tested their approach on five-minute EUR/USD exchange market data and came to the conclusion that the use of their a-priory robustness criterion improves both robustness and profitability. Despite of this, they did not get to find profitable strategies in their experiments.

We could mention a related study Oesch and Maringer [25] where these authors use GE to develop a high frequency trading system that exploits volume inefficiencies at the bid–ask spread. Their experimental results show that the system identifies strategies that are both profitable and robust.

Martín et al. [26] introduced a GE-based ensemble approach that includes a voting mechanism with an inertia component that balances a certain degree to adaptability to structural change while limiting the number of trading signals. The results, based on S&P500 data, suggest that this strategy beats the traditional static approach that relies on a single model for the whole period and ensembles that implement more standard voting systems.

The approach that we present in this work poses a number of advantages vs other alternatives discussed above. The most important of them are the ability to adapt automatically to structural changes, and the fact that it is done in a way that limits the overtrading that often drags down the performance of other algorithms.

Finally, it is worth mentioning that the applications of GE are by no means limited to this field. This technique might potentially be used to tackle other problems related to economics and management, and its potential should be explored in domains like blockchain in banking [27]-[28] or industrial diagnosis [29].

## III. Traditional Static Approach

Grammatical evolution, a metaheuristic closely related to genetic programming developed by Ryan et al. [3], encodes individuals as strings of integers that are mapped to programs by means of context-free grammars. Like other techniques within the framework of evolutionary computation, it is a stochastic population-based approach that refines solutions in an iterative way by means of the application of a number of operators (selection, crossover, mutation etc.) according to a basic loop.

While genotypes take the form to strings of integers, phenotypes are structured as Lisp-style functional trees. The connection between these two elements is managed by user-specified grammars that describe the core elements of the programs: terminals, non-terminals and the associated lexis and syntax. It is worth noting that these descriptions, usually in Backus-Naur form (BNF), often incorporate domain knowledge and constrain the search space. The use of grammars offers an important advantage over standard genetic programming, like simultaneously enforcing closure and allowing different data types.

The process starts with the initialization of the population. This requires the generation of a as many vectors of integers as individuals. The vectors are then mapped to terminal and non-terminal elements according to the grammar. The initialization gradually generates functional trees (or their s-string equivalents) until either all the non-terminal nodes get the required inputs, or all the initial contents of the vectors get used, case in which these are expanded with as many additional elements as necessary.

The range of trading rules that can be generated, is determined by the selection of the set of terminal and non-terminal elements, hence its importance. For the purposes of this study we basically relied on a previous study by Lohpetch and Corne [7]. The only difference was that we used daily data instead of monthly information. This choice aligns the study with most of the ones mentioned in the previous section. The set of non-terminal nodes included two relational operators (>and <) and three logical ones (*And*, *Or* and *Not*).

Regarding the terminal elements, we considered a number of technical indicators and the most important basic daily index prices, including *Opening* and *Closing*, *Maximun* and *Minimum*). Simple moving averages, another popular indicator, were included for 2, 3, 5 and 10-months (*MA2*, *MA3*, *MA5* and *MA10*). We included the 3-month as a momentum oscillator, and 12-month Rate of Change indicator (*Roc3* and *Roc12*) and the 3-month rolling minima and maxima (*Mx1*, *Mx2*, *Min1* and *Min2*) as short-term price resistance indicators. Finally, we added two trend-line indicators, upper and lower resistance lines (*UR* and *LR*) to characterize the speed and direction of price changes.

Generating syntactically correct investment rules based on these elements requires the definition of grammars. As we mentioned, these specify which operators and terminals are acceptable as input arguments, together with their appropriate outputs. In this study we also follow the one introduced in [7], which is detailed in Table I in BNF-form. The application of this grammar results in the obtainment of conditional rules that were interpreted as recommendations for being invested, "1", or in cash, "0", depending on market conditions.

Fig. 1 illustrates the structure of these trading rules showing one that could be potentially generated using this approach. In this case, the rule would trigger a recommendation to be invested if either the maximum trading price is larger than the two-month rolling maximum, the two-month rolling maximum is larger than the lower trend line indicator or the 5-month moving average is larger than the 2-month one, or both.

TABLE I. Grammar Used to Define Trading Rules

| Nº | Modulus | Grammar Rule |
|---|---|---|
| 1 | 1 | <Rule> ::= <bool> |
| 2 | 5 | <Bool> ::= (**And** <bool><bool>)\| (**Or** <bool><bool>) |
| | | <Bool> ::= (**Not** <bool>) |
| | | <Bool> ::= (><exp><exp>) \| (<<exp><exp>) |
| 3 | 16 | <Exp>::= (**Opening**)\|(**Closing**)\|(**Maximun**)\|(**Minimum**) |
| | | <Exp> ::= (**Me2**) \| (**Me3**) \| (**Me5**) \| (**Me10**) |
| | | <Exp> ::= (**Roc3**)\| (**Roc12**) |
| | | <Exp> ::= (**Max1**) \| (**Max2**) \| (**Min1**) \| (**Min2**) |
| | | <Exp> ::= (**UR**) \| (**LR**) |

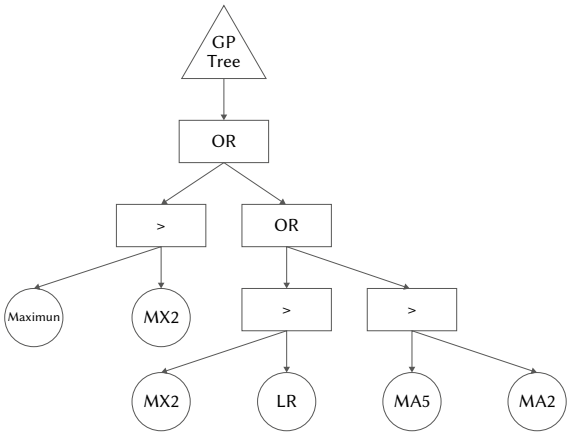



Fig. 1. Example of trading rule represented as a tree. This would be equivalent to the Common Lisp S-expression (Or (>Maximun Mx2) (Or (>Mx2 LR) (>MA5 MA2))).

Once we have discussed rule structure, we will consider evaluation. The quality of trading rules was assessed in terms of net profit or loss. Return, $r$, was characterized as the sum of returns minus transaction costs using continuous compounding like in [4] and [30]. Hence, the fitness function could be defined formally as

$$r = \sum_{t=1}^{T} r_t \cdot I_b(t) + \sum_{t=1}^{T} r_f(t) \cdot I_s(t) + n \cdot ln\left(\frac{1-c}{1+c}\right) \quad (1)$$

where $r_t = ln(P_t) - ln(P_{t-1})$ represents the return on the index computed as its price difference between time $t-1$ and $t$; $Ib(t)$ is a dichotomous variable that is equal to one in the periods where the rule recommends being invested and zero in the rest; $I_s(t)$ is $1 - b(t)$, and $r_f(t)$ is the risk-free rate of return for one period of time prevailing at time $t$.

Regarding the last term of the expression, it is an estimate for the transaction costs resulting from the purchases and sales derived from moving from a recommendation to be invested in the market to hold a cash position, and vice-versa. Here $n$ represents number of transactions and $c$ the one-way transaction cost as a fraction of price (in the experimental analysis it was set at 0.25%).

This process is illustrated in Fig. 2, where we can see the investment positions recommended by an investment rule on the S&P 500 over 250 trading days. There, we show the behavior of the index together with an overlay using a thicker line that shows the return accumulation process. The strategy tracks the market, and therefore its performance, whenever the evaluation of the rule for the day returns "1", and falls down to the bottom in case the recommendation for the day is being out of the market. Initially the rule recommends being in cash for about a month, accruing the risk-free interest rate, and then investing in the index for 4 days.

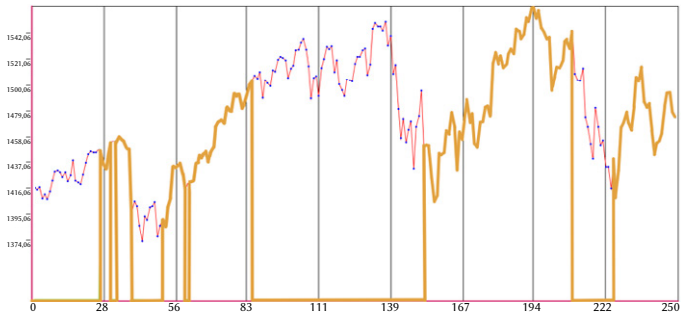



Fig. 2. Example of trading rule performance evaluation. Thicker line illustrates the recommended investment position and return accumulation for a trading rule over the period.

In this case, the evolved trading rule recommends moving in and out of the market several times. In the process, it anticipates two important market corrections, and prevents losses staying in cash over the period from approximately day 84 to 151 and between trading days 208 and 224.

The main loop of GE requires the use genetic operators to drive the iterative improvement process at the core of the metaheuristic. These are very similar to the ones used in the more popular GP, making the range of alternatives very wide. In this study we relied on standard ones: tournament selection, single-point crossover and uniform mutation. Given that the use of these operators often results in a significant number of malformed individuals, we implemented two standard repair mechanisms. The first one, duplication, was used whenever the new individual happened to be too short. This strategy replicates a portion of the individual, selected randomly, and extends the genome up to the required size. The second, truncation, was used in the opposite scenario. It disregards the final elements of the sequence of integers that are unnecessary.

Finally, the core setup used in this study implements non-parametric parsimony pressure, a mechanism that punishes complexity selecting the simplest rule whenever there is a tie in the selection operator. This limit bloating and, as a result, enhances performance, reducing overfitting. In addition to that, this offers the advantage of improving interpretability, a key aspect in this domain.

## IV. Dynamic Approach

The adaptive approach we suggest involves a dynamic selection mechanism that commutes between an active rule and a candidate one, optimized for the most recent available market data.

The system relies on the use of sliding windows. Given a window size, defined by a constant $w$, the process starts selecting the $w$ most recent data points to evolve a trading rule using GE. Once a rule is obtained, it can be used to generate investment recommendations for future periods. If we do this only once, and we use the resulting strategy over the whole test period, we obtain the standard non-adaptive approach often found in the literature. We will label it *Static*.

On the opposite end of the spectrum, we might repeat the process to generate rules that would only be used once for a single time-step. That is, if we moved the sliding window one time-step at a time, we would obtain overlapping training samples that would differ in a single element. The new one would add the most recent data point, and drop the oldest one. If we considered only the most recent rule to provide an investment recommendation for the next period, and we iterated, we would obtain a very adaptive strategy that we will call *Naif*. These rules could also be combined into ensembles so that the recommendations of the $e$ most recent rules for any specific period could be combined into a single one either using a simple majority voting approach, *Majority*, or a weighted voting mechanism, *Weighted*.

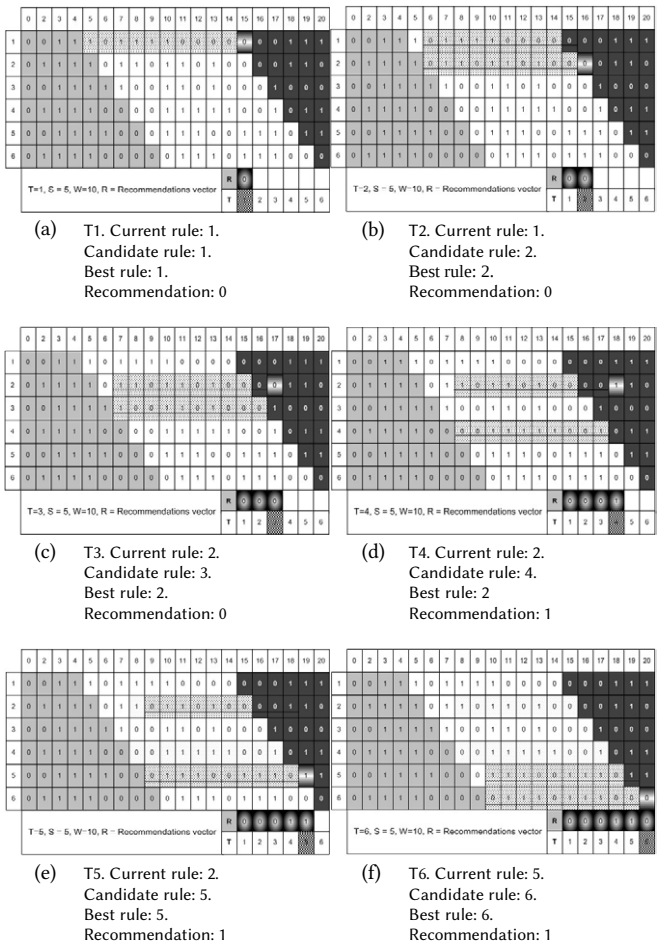
The solution that we introduce intends to achieve a balance between the need to adapt to structural change and to limit the impact of over-trading. The starting point would be the *Naif* approach, which evolves a trading rule for each time step using a sliding window with a fix window size $w$. Given a time $t$, a trading rule is evolved using GE in the period between $t - w$ and $t - 1$. Given that it would be the only alternative, it would be considered the best rule and therefore, it be used to generate the first investment recommendation. At $t$+1, a candidate rule would be evolved based on the period from $t$–$w$+1 to $t$. At that point, the current best rule and the new candidate rule would be compared. The best one would get the status of current best rule and therefore, would be used to generate the required investment recommendation for $t + 1$.

The process of updating the best investment rule comparing the performance of the current one vs the candidate rule based on the most recent $w$ data points is repeated at every time-step. As a result, we obtain a dynamic investment strategy. It is worth noting that once the current best rule is replaced by a new one, it is lost. The process can only move forward.

Rule comparison is a key aspect that requires clarification. It is made on the basis of investment performance on a common evaluation period that will always be the training period for the new candidate rule (the most recent $w$ data points). This also means that, given period overlaps, from the point of view of current best rule, performance will always be based on the evaluation of investment recommendations on a period that includes test data to some degree. If the current best performing model had been updated in the previous period (and, therefore, had generated recommendations for one period only) and we use it as reference point, we would have $w - 1$ recommendations on training data and 1 recommendation on test data. On the other end of the spectrum, if the best current model had maintained its status for $w$ or more consecutive time periods, we would consider the whole $w$ recommendations on test data. Once again, we would like to emphasize that this explanation uses the current best-performing model as the reference point. As we mentioned, as long as the new candidate rule is concerned, the two rules are compared based on performance on its training sample.

This process is illustrated with an example in Fig. 3. Panel 2a describes the selection of a first inversion rule, obtained with GE, on a training sample of $w = 10$ periods, $t_5 - t_{14}$ (row 1, in dotted black and white). Being the first rule, it is chosen as an active rule and will generate the recommendation. Once the first recommendation is obtained, step T1, "0", $t_{15}$ (in black and white), the sliding window moves one period to the right and the second rule is generated, panel 2b, based on the time period $t_6 - t_{15}$, (row 2, in dotted black and white). The performance of this candidate rule is compared with the performance of the best rule in progress, rule 1, in the same period (marked with black lines). That means we would consider 9 time periods that were part of the training sample for the active rule, $t_6 - t_{14}$ (row, 2, in dotted black and white), and one of the test sample, $t_{15}$ (in black). Based on that, the candidate rule, 2, would take over the role of current best rule, and its recommendation, "0", would be used as the output of the system for T2, the second time step of the test period $t_{16}$ (in black and white) and the sixteenth element of the sequence.

In T3, in 2c, the best current rule would be the second one, and the new candidate would be obtained based on the period from $t_7$ to $t_{16}$ (row 3, in dotted black and white). In the example, the comparison in this period would favor rule 2 and, as a result, the candidate rule, 3, would be discarded and, once again, the output of the system would mirror the recommendation of the second rule, "0", to stay out of the market. The dynamics of the process in test periods 4-6, represented in 2d - 2f are very similar. As the sliding window moves, it generates new candidate rules that are compared with the best alternative at that time, updating them accordingly.



(a) T1. Current rule: 1. Candidate rule: 1. Best rule: 1. Recommendation: 0

(b) T2. Current rule: 1. Candidate rule: 2. Best rule: 2. Recommendation: 0

(c) T3. Current rule: 2. Candidate rule: 3. Best rule: 2. Recommendation: 0

(d) T4. Current rule: 2. Candidate rule: 4. Best rule: 2 Recommendation: 1

(e) T5. Current rule: 2. Candidate rule: 5. Best rule: 5. Recommendation: 1

(f) T6. Current rule: 5. Candidate rule: 6. Best rule: 6. Recommendation: 1

Fig. 3. Illustration of algorithm behavior. Rule selection for test periods 1 to 6. Training samples in darker color. Test samples in lighter color. Investment recommendations by rule: "1" stay in the market "0" stay in cash. Basis of rule comparison in rectangles.

## V. Experimental Analysis

This section describes the experimental work used to evaluate the dynamic approach introduced in the previous section. We start discussing the main aspects of the experimental setup, including data set, experimental protocol and parameterization, to then analyze the results.

### A. Experimental Setup

The performance of the *Dynamic* approach was assessed comparing its returns vs. the four comparable strategies described in 4: *Static*, *Naif*, and the two ensemble-based alternatives, *Majority* and *Weighted*, to time the Standard & Poor's 500 index.

The sample covered 13 years of daily data starting from the beginning of 2005 to the end of 2017. In addition to the index information, we also needed the historical daily short-term risk-free interest rates of return over the same periods. The former was obtained from the commercial provider Datastream, while interest rates were downloaded from the Federal Reserve Bank of Atlanta.

The main data set was divided in two. The first sample, which covered from 2005 to the end of 2012 was used to run the exploratory experiments for parametrization purposes while the rest was kept for testing. Once the parameters were chosen, the final performance was evaluated on an annual basis in the period from the beginning of

2013 to the end of 2017. Given the evaluation period, we also used the last portion of the first sample to train some of the models that were compared in the test sample.

While the decision to break down the comparative analysis in 5 sub-periods instead of using only one makes no difference in the way the models are trained, we understand that this kind of evaluation provides relevant information on the evolution of reliability over time. Given that GE is a stochastic method we repeated the experiments 30 times.

The exploratory analysis resulted in the selection of population size of 500 individuals, which evolved over 50 generations carrying over the best individual of every generation to the next one. The population was initialized using geometric series with a minimum initial complexity of 5 and a probability of growth of 0.85.

Regarding the main operators, we used simple tournament with a size of 2, and applied the following to those selected: one-point crossover with probability of 0.85, duplication with a probability of 0.05, and uniform mutation with a probability of 0.1. The latter randomly modified genes within a specified range (-128, 127) with a probability of 0.05, allowing a circular wrapping of gene vector up to 16 times before discarding the individual as invalid.

In order to improve diversity and avoid premature convergence, both during the initialization process of the algorithm and during the genetic mutation operation, if the same individual appeared more than once in the population, we tried up to 100 attempts to replace it with a new one.

The size of the training window was set at three years (753 sessions), while the number of sessions in the test period was fixed at one year (251 sessions).

In relation to the ensembles, the number of rules, *e*, considered to generate the recommendations was set to 5 and the weights of the *Weighted* approach were set to [0.05,0.1,0.15,0.25,0.45]. Here, the last elements of the vector make reference to the emphasis given to the rules based on the most recent information. That is, the vote of the most recent rule would have significantly more importance than the rest.

### B. Results

The main results of the experiments are summarized in Table II. There, we report the most important descriptive statistics for the net returns over 30 experiments. The table details the performance of the *Dynamic* approach plus the two benchmarks that represent the opposite extremes in terms of adaptation to structural change, *Static* and *Naif*, plus the two ensembles, *Majority* and *Weighted*, described before.

As we can see, the *Dynamic* strategy provided the best average performance. If we focus on average yearly return over the whole 5-year period, it offered 10.71% net return. Meanwhile, the performance of the most competitive alternative, the *Static* approach, was 2.54%. This result, though significantly poorer, was still better than the one obtained by the ensembles and, specially, the *Naif* one, which resulted in an average yearly net loss of 3.96%. If we consider reliability, the *dynamic* strategy also yielded more consistency, as the average of the yearly return variances was very low compared to the one obtained using the *Static* approach.

Once we analyze the results year by year, we can see that *Dynamic* clearly dominates *Static*, *Naif*, *Majority* and *Weighted* regardless of market conditions. Although the rank of the four alternatives changes between them over time, *Dynamic* consistently outperforms the four of them.

The significance of the reported mean performance differences vs. *Dynamic* was formally tested. The process followed to that end started with the assessment of the normality of the distribution of returns using Kolmogorov-Smirnov test with the Lilliefors correction [31]. Whenever the normality of the results was rejected, we relied on non-parametric Wilcoxon' test [32]. Conversely, in case that it could not be rejected, we used Levene's homoskedasticity test [33]. At that point, depending on the result, we employed either a t-test [34] or Welch's [35]. According to this protocol, all the differences with the exception of one were significant at 1%. The observed performance differences might be explained by two main reasons: better market timing and better control of transaction costs. The latter aspect was analyzed tracking the number of purchase and sale orders generated by the three methods. This information is reported in Table III.

TABLE II. Net Return. Main Descriptive Statistics Over 30 Runs. Test Results

| | Strategy | Mean | | Median | Var. | Max. | Min. |
|---|---|---|---|---|---|---|---|
| 2013 | Dynamic | 0.2161 | | 0.2318 | 0.0017 | 0.2543 | 0.1311 |
| | Static | 0.0381 | * | 0.0409 | 0.0001 | 0.0715 | 0.0051 |
| | Naif | 0.1318 | ** | 0.1318 | 0.0002 | 0.1520 | 0.0921 |
| | Majority | 0.1367 | ** | 0.1363 | 0.0002 | 0.1557 | 0.0964 |
| | Weighted | 0.1372 | ** | 0.1387 | 0.0001 | 0.1582 | 0.1088 |
| 2014 | Dynamic | 0.0922 | | 0.0994 | 0.0002 | 0.1042 | 0.0522 |
| | Static | 0.0606 | ** | 0.0827 | 0.0015 | 0.1042 | 0.0079 |
| | Naif | 0.0559 | ** | 0.0567 | 0.0002 | 0.0817 | 0.0257 |
| | Majority | 0.0852 | ** | 0.0876 | 0.0001 | 0.0969 | 0.0471 |
| | Weighted | 0.0782 | ** | 0.0784 | 0.0001 | 0.0961 | 0.0548 |
| 2015 | Dynamic | -0.0152 | | -0.0123 | 0.0002 | -0.0123 | -0.0935 |
| | Static | -0.0160 | * | -0.0123 | 0.0001 | -0.0123 | -0.0601 |
| | Majority | -0.0903 | ** | -0.0817 | 0.0017 | -0.0303 | -0.1738 |
| | Weighted | -0.1266 | ** | -0.1270 | 0.0016 | -0.0488 | -0.1958 |
| 2016 | Dynamic | 0.0758 | | 0.0889 | 0.0012 | 0.0889 | -0.0217 |
| | Static | 0.0241 | ** | 0.0043 | 0.0016 | 0.0889 | -0.0223 |
| | Naif | -0.1411 | ** | -0.1440 | 0.0014 | -0.0335 | -0.2193 |
| | Majority | -0.0448 | ** | -0.0383 | 0.0012 | 0.0075 | -0.1231 |
| | Weighted | -0.0828 | ** | -0.0778 | 0.0023 | 0.0043 | -0.1922 |
| 2017 | Dynamic | 0.1668 | | 0.1668 | 0.0000 | 0.1668 | 0.1668 |
| | Static | 0.0204 | ** | 0.0092 | 0.0016 | 0.1668 | 0.0092 |
| | Naif | -0.0718 | ** | -0.0755 | 0.0011 | 0.0295 | -0.1316 |
| | Majority | 0.0176 | ** | 0.0294 | 0.0022 | 0.1094 | -0.0753 |
| | Weighted | 0.0025 | ** | 0.0057 | 0.0023 | 0.0943 | -0.1136 |
| Mean | Dynamic | 0.1071 | | 0.1149 | 0.00066 | 0.1204 | 0.0470 |
| | Static | 0.0254 | | 0.0250 | 0.00099 | 0.0838 | -0.0120 |
| | Naif | -0.0396 | | -0.0419 | 0.00104 | 0.0311 | -0.0985 |
| | Majority | 0.0209 | | 0.0267 | 0.0011 | 0.0678 | -0.0458 |
| | Weighted | 0.0017 | | 0.0036 | 0.0013 | 0.0608 | -0.0676 |

** Significant vs. Dynamic at 1%

While it is clear that the *Naif* strategy was, by far, the most active one, the rank of the rest is not stable. The *Dynamic* strategy traded less often than the rest 3 out of the five years. The performance of the *Naif* approach was therefore severely undermined, as it is apparent once we analyze performance in gross terms.

Table IV is similar to Table II. The difference is that it represents gross returns and, therefore, the performance has not been adjusted for transaction costs. If we consider average yearly performance over the 2013-2017 period, we see that the 11.56% return offered by the *Dynamic* approach beats both the *Naif* and the *Static* one, with 7.20% and 3.72% respectively. That is also the case for the ensembles, as the one based on simple majority obtained an average gross return of 6.66%, very similar to the one provided by the one based on weighted voting with 6.63%.

TABLE III. Number of Transactions

| | Strategy | Mean | Median | Var. | Min. | Max. |
|---|---|---|---|---|---|---|
| 2013 | Dynamic | 4.93 | 4 | 6.82 | 2 | 10 |
| | Static | 7.20 | 4 | 48.17 | 4 | 32 |
| | Naif | 14.00 | 14 | 16.00 | 4 | 20 |
| | Majority | 5.80 | 6 | 3.96 | 4 | 10 |
| | Weighted | 5.07 | 4 | 1.86 | 4 | 10 |
| 2014 | Dynamic | 4.40 | 4 | 5.63 | 2 | 10 |
| | Static | 7.07 | 4 | 25.31 | 2 | 14 |
| | Naif | 14.73 | 14 | 11.72 | 8 | 22 |
| | Majority | 8.00 | 8 | 1.66 | 6 | 10 |
| | Weighted | 9.13 | 8 | 4.6 | 6 | 14 |
| 2015 | Dynamic | 2.40 | 2 | 3.42 | 2 | 12 |
| | Static | 2.40 | 2 | 1.21 | 2 | 6 |
| | Naif | 42.40 | 42 | 24.94 | 36 | 56 |
| | Majority | 17.00 | 18 | 10.69 | 10 | 26 |
| | Weighted | 24.33 | 24 | 21.26 | 18 | 40 |
| 2016 | Dynamic | 2.73 | 2 | 8.69 | 2 | 18 |
| | Static | 5.53 | 4 | 13.43 | 2 | 18 |
| | Naif | 72.47 | 72 | 58.40 | 62 | 92 |
| | Majority | 29.33 | 28 | 19.95 | 18 | 40 |
| | Weighted | 43.13 | 42 | 36.33 | 28 | 58 |
| 2017 | Dynamic | 2.00 | 2 | 0.00 | 2 | 2 |
| | Static | 1.40 | 0 | 13.28 | 0 | 12 |
| | Naif | 79.40 | 81 | 102.39 | 52 | 100 |
| | Majority | 31.33 | 32 | 35.40 | 20 | 44 |
| | Weighted | 45.93 | 48 | 111.86 | 30 | 78 |
| Mean | Dynamic | 3.29 | 2.80 | 4.91 | 2.00 | 10.40 |
| | Static | 4.72 | 2.80 | 20.28 | 2.00 | 16.40 |
| | Naif | 44.60 | 44.60 | 42.69 | 42.00 | 48.40 |
| | Majority | 18.29 | 18.40 | 17.50 | 11.60 | 26.00 |
| | Weighted | 25.52 | 25.20 | 35.18 | 17.20 | 40.00 |

TABLE IV. Gross Return. Main Descriptive Statistics Over 30 Runs. Test Results

| | Strategy | Mean | | Median | Var. | Max. | Min. |
|---|---|---|---|---|---|---|---|
| 2013 | Dynamic | 0.2297 | | 0.2531 | 0.0014 | 0.2593 | 0.1561 |
| | Static | 0.0561 | ** | 0.0509 | 0.0004 | 0.1215 | 0.0193 |
| | Naif | 0.1661 | ** | 0.1654 | 0.0003 | 0.2215 | 0.1409 |
| | Majority | 0.1512 | ** | 0.1519 | 0.0001 | 0.1676 | 0.1214 |
| | Weighted | 0.1540 | ** | 0.1562 | 0.0001 | 0.1732 | 0.1338 |
| 2014 | Dynamic | 0.1032 | | 0.1092 | 0.0001 | 0.1141 | 0.0721 |
| | Static | 0.0783 | ** | 0.0927 | 0.0007 | 0.1092 | 0.0429 |
| | Naif | 0.0915 | ** | 0.0890 | 0.0001 | 0.1165 | 0.0716 |
| | Majority | 0.1052 | | 0.1074 | 0.0001 | 0.1184 | 0.0671 |
| | Weighted | 0.1010 | | 0.0987 | 0.0001 | 0.1161 | 0.0798 |
| 2015 | Dynamic | -0.0092 | | -0.0073 | 0.0001 | -0.0073 | -0.0635 |
| | Static | -0.0100 | | -0.0073 | 0.0001 | -0.0073 | -0.0451 |
| | Naif | -0.0627 | ** | -0.0627 | 0.0024 | 0.0576 | -0.1379 |
| | Majority | -0.0478 | ** | -0.0415 | 0.0014 | 0.0084 | -0.1288 |
| | Weighted | -0.0658 | ** | -0.0692 | 0.0014 | 0.0112 | -0.1361 |
| 2016 | Dynamic | 0.0826 | | 0.0939 | 0.0008 | 0.0939 | -0.0042 |
| | Static | 0.0379 | ** | 0.0293 | 0.0015 | 0.0952 | -0.0073 |
| | Naif | 0.0357 | ** | 0.0409 | 0.0018 | 0.1346 | -0.0282 |
| | Majority | 0.0285 | ** | 0.0355 | 0.0010 | 0.0775 | 0.0481 |
| | Weighted | 0.0250 | ** | 0.0275 | 0.0019 | 0.1243 | -0.0497 |
| 2017 | Dynamic | 0.1718 | | 0.1718 | 0.0000 | 0.1718 | 0.1718 |
| | Static | 0.0239 | ** | 0.0092 | 0.0017 | 0.1718 | 0.0092 |
| | Naif | 0.1296 | ** | 0.1290 | 0.0013 | 0.1968 | 0.0591 |
| | Majority | 0.0959 | ** | 0.0294 | 0.0014 | 0.1694 | 0.0323 |
| | Weighted | 0.1174 | ** | 0.1110 | 0.0014 | 0.1922 | 0.0479 |
| Mean | Dynamic | 0.1156 | | 0.1241 | 0.00048 | 0.1264 | 0.0665 |
| | Static | 0.0372 | | 0.0350 | 0.00088 | 0.0981 | 0.0038 |
| | Naif | 0.0720 | | 0.0723 | 0.00119 | 0.1454 | 0.0211 |
| | Majority | 0.0666 | | 0.0565 | 0.0008 | 0.0280 | 0.1082 |
| | Weighted | 0.0663 | | 0.0648 | 0.0010 | 0.0151 | 0.1234 |

** Significant vs. Dynamic at 1% * Significant vs. Dynamic at 5%

The breakdown by year shows that *Naif* profited much than *Static* in the periods where there was more to be gained. Losses in bad years were also mitigated to a very large extent. It is worth mentioning that disregarding transaction costs turned the 2016 major losses for the *Naif* strategy into profits. Conversely, the *Static* approach evolved a large proportion of trading rules that did not provide any trading signals, hence making the difference between net a gross performance negligible. The ensembles generally performed in line with the *Naif* in gross terms, but they obtained very good results in 2014, where *Majority* got to beat *Dynamic*. Having said that, the difference was small and not statistically significant.

As we also observed when we analyzed net returns, the *Dynamic* approach introduced in this study seems to be the most reliable one in terms gross performance. The average of yearly return variances was around half of the of the second most stable approach.

The fact that the *Dynamic* approach also offered such good results in gross terms indicate that the dominance that we observed in net returns can be explained by a both combination a combination of adaptability to structural change and limited transaction costs. The *Naif* strategy identified and exploited small structural changes, but the excessive trading caused by constant replacements of investment rules increased transaction costs to the point of making flexibility counterproductive. This is likely to be caused by the fact that trading rules identified using GE are implicitly optimized to limit the number of signals, as they are penalized in the fitness function. Once there is a constant change, the strategy that is used in practice is not none of the optimized ones and, therefore, the implicit control mechanism for transaction costs is likely to be affected. The *Static* approach does not offer the flexibility of the rest but, if structural change is limited over the period of use, the trade-off of limited transaction costs vs the loss of accuracy over time could still lead to good results.

The experimental results support the importance of using dynamic approaches like the described one. The advantages are not limited to the ability to generate relevant trading signals in a dynamic environment, which is clearly an important aspect, but also the possibility of doing it at the same time that it controls transaction costs. The proposed strategy *Dynamic* offers a good compromise between two often conflicting objectives: offering flexibility to adapt to structural changes and limiting the number of orders.

## VI. Summary and Conclusions

The development of investment rules using grammatical evolution often entails obtaining a single rule based on a training period, which is then used to generate recommendations over time. The use of sliding windows improves the adaptability to the structural changes that prevail in financial series but tends to result in excessive transaction costs. Therefore, it is necessary to find alternatives that offer a balance

between flexibility and transaction expenses.

In this study we improve the standard approach introducing new solution that involves a dynamic selection mechanism, which switches between an active rule and a candidate one optimized for the most recent market data available. The process also includes a hysteresis component that reduces the risk of overtrading.

The approach was benchmarked against four alternatives based on the same core algorithm over a period of five years. The alternatives included: the standard static approach *Static,* a solution that updates constantly the decision rule, *Naif*, and two ensemble-based solutions that differ in the voting mechanism that they implement, *Majority* and *Weighted.*

The results obtained support the superiority of the new solution both in terms of return and reliability, followed by the *Static* approach.

The analysis of the impact of transaction costs on profitability highlights the importance of limiting overtrading, since there is a clear inverse relationship between the number of purchase and sale orders and performance. The *Naif* approach trades much more often than the rest, and commissions drag down its returns to a very large extent. It is worth noting that controlling this aspect seems to be a key success factor of the *Dynamic* alternative, but not the only one.

These findings bring out the importance of holding a balance between the importance to adapt to market structural changes and the risk of updating constantly recommendation models that are implicitly optimized for the longer term. The results support the new approach as a mechanism capable of maintaining the balance sought between these two contradictory goals.

Future lines of work might include replicating the study with other assets or financial indices; testing the approach with genetic programming; extending the grammar to analyze the impact on the results, or exploring the possibility of updating dynamically the set of terminals and nodes to make sure that the building blocks of the rules remain relevant over time.

## Acknowledgment

The authors would like to acknowledge the financial support of the Spanish Ministry of Science, Innovation and Universities under grant PGC2018-096849-B-I00 (MCFin).

This work has been supported by the Madrid Government (Comunidad de Madrid-Spain) under the Multiannual Agreement with UC3M in the line of Excellence of University Professors (EPUC3MXX), and in the context of the V PRICIT (Regional Programme of Research and Technological Innovation).

## References

[1] F. Allen, R. Karjalainen, "Using genetic algorithms to find technical trading rules", *Journal of Financial Economics* vol. 51, no. 2, pp. 245–271, 1999.

[2] J.R. Koza, "Genetic programming as a means for programming computers by natural selection", *Statistics and Computing*, vol. 4, no. 2, pp. 87–112 1994.

[3] C. Ryan, J. Collins, M. O'Neil, "Grammatical Evolution: Evolving Programs for an Arbitrary Language", in *Proc. of the First European Workshop on GP*. Springer Berlin Heidelberg pp. 83–96, 1998.

[4] C. Setzkorn, L. Dipietro, R. Purshouse, "Evolving Rule-Based Trading Systems", Technical Report ULCS-02-005, Department of Computer Science, University of Liverpool, 2002.

[5] C.J. Neely, "Risk-adjusted, ex ante, optimal technical trading rules in equity markets", *International Review of Economics & Finance*, vol. 12, pp 69–87, 2003.

[6] N. Navet, S.H. Chen, "On predictability and profitability: Would GP induced trading rules be sensitive to the observed entropy of time series?", *Studies in Computational Intelligence*, vol. 100, pp. 197–210, 2008.

[7] D. Lohpetch, D. Corne, "Discovering effective technical trading rules with genetic programming: Towards robustly outperforming buy-and-hold" in *Proceedings of the World Congress on Nature and Biologically Inspired Computing, NABIC 2009*, pp. 439–444.

[8] D. Lohpetch, D. Corne, "Discovering effective technical trading rules with genetic programming: Towards robustly outperforming buy-and-hold", in *Lecture Notes in Computer Science*, vol. 6025, 2010, pp. 171–181.

[9] J. How, M. Ling, P. Verhoeven, "Does size matter? A genetic programming approach to technical trading" *Quantitative Finance*, vol. 10, no. 2, pp. 131–140, 2010.

[10] A. Esfahanipour, S. Mousavi, "A genetic programming model to generate risk-adjusted technical trading rules in stock markets", *Expert Systems with Applications* vol. 38, no. 7, pp. 8438–8445, 2011.

[11] V. Manahov, R. Hudson, H. Hoque, "Return predictability and the 'wisdom of crowds': Genetic Programming trading algorithms, the Marginal Trader Hypothesis and the Hayek Hypothesis", *Journal of International Financial Markets, Institutions and Money*, vol. 37, pp. 85–98, 2015.

[12] S.H. Chen, T.W. Kuo, K.M. Hoi, Genetic Programming and Financial Trading: How Much About "What We Know" In: Zopounidis C., Doumpos M., Pardalos P.M. (eds) *Handbook of Financial Engineering. Springer Optimization and Its Applications,* vol. 18 (Springer US, Boston, MA, 2008), pp. 99–154.

[13] E.A. Gerlein, M. McGinnity, A. Belatreche, S. Coleman, "Evaluating ma- chine learning classification for financial trading: An empirical approach", *Expert Systems with Applications*, vol. 54, pp. 193–207, 2016.

[14] R. Ray, P. Khandelwal, B. Baranidharan, "A survey on stock market prediction using artificial intelligence techniques", in *2018 International Conference on Smart Systems and Inventive Technology (ICSSIT),* pp. 594–598.

[15] T.L. Meng, M. Khushi, "Reinforcement learning in financial markets", *Data*, vol. 4, no. 3, 110, pp. 1–17, 2019.

[16] A. Brabazon, "Grammatical Evolution in Finance and Economics: A Survey", in Ryan C., O'Neill M., Collins J. (eds) *Handbook of Grammatical Evolution.* Springer, pp. 263–288, 2018.

[17] C. Martín, D. Quintana, P. Isasi, "Evolution of trading strategies with flexible structures: A configuration comparison". *Neurocomputing*, vol. 331, pp. 242–262, 2019.

[18] A. Brabazon, M. O'Neill, "Evolving technical trading rules for spot foreign-exchange markets using grammatical evolution", *Computational Management Science*, vol. 1, pp. 311–327, 2004.

[19] I. Dempsey, M. O'Neill, A. Brabazon, "Evolving technical trading rules for spot foreign-exchange markets using grammatical evolution", *GECCO 2004 Workshop Proceedings* pp. 9137–9142, 2004.

[20] I. Contreras, J.I. Hidalgo, L. Nunez-Letamendia, "Evolving technical trading rules for spot foreign- exchange markets using grammatical evolution", in *Applications of Evolutionary Computing, EvoApplications 2013: EvoCOMNET, EvoCOMPLEX, EvoENERGY, EvoFIN, EvoGAMES, EvoIASP, EvoINDUSTRY, EvoNUM, EvoPAR, EvoRISK, EvoROBOT, EvoSTOC*, vol. 7835 (Springer, Berlin, Heidelberg, 2013), pp. 244–253.

[21] I. Contreras, J.I. Hidalgo, L. Núñez-Letamendia, "A GA Combining Technical and Fundamental Analysis for Trading the Stock Market", in *A GA combining technical and fundamental analysis for trading the stock market* (Springer, Berlin, Heidelberg, 2012), pp. 174–183.

[22] I. Contreras, J.I. Hidalgo, L. Núñez-Letamendia, "Evolving technical trading rules for spot foreign-exchange markets using grammatical evolution", *Journal of Intelligent and Fuzzy Systems*, vol. 32, no. 3, pp. 2461–2475, 2017.

[23] I. Contreras, J.I. Hidalgo, L. Nuñez-Letamendia, J.M. Velasco, "A metagrammatical evolutionary process for portfolio selection and trading", *Genetic Programming and Evolvable Machines*, vol. 18, no. 4, pp. 411–431, 2017.

[24] H. Schmidbauer, A. Rösch, T. Sezer, V.S. Tunaliog˘lu, "Robust trading rule selection and forecasting accuracy", *Journal of Systems Science and Complexity*, vol. 27, no. 1, pp. 169–180, 2014.

[25] C. Oesch, D. Maringer, "Robust trading rule selection and forecasting accuracy", *Quantitative Finance,* vol. 17, no. 5, pp. 717–727, 2017.

[26] C. Martín, D. Quintana, P. Isasi, "Grammatical evolution-based ensembles for algorithmic trading", *Applied Soft Computing*, vol. 84, 105713, pp. 1–10, 2019.

[27] R. Arjun, K. R, Suprabha, "Innovation and Challenges of Blockchain in Banking: A Scientometric View". *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 3, pp. 7–14, 2020. 10.9781/ijimai.2020.03.004.

[28] A.T. Chatfield, C. Reddick, "Blockchain Investment Decision Making in Central Banks: A Status Quo Bias Theory Perspective", in *Proceedings of Americas Conference on Information Systems AMICS 2019 Proceedings* pp. 1–10.

[29] F.Z. Benkaddour, N. Taghezout, F.Z. Kaddour-Ahmed, I.-A. Hammadi, "An adapted approach for user profiling in a recommendation system: Application to industrial diagnosis", *International Journal of Interactive Multimedia and Artificial Intelligence*, vol 5, no. 3, pp. 118–130, 2018.

[30] L. Becker, M. Seshadri, "GP-evolved technical trading rules can outperform buy and hold", in *Proceedings of the Sixth International Conference on Computational Intelligence and Natural Computing*, Embassy Suites Hotel and Conference Center, Cary, North Carolina USA, September 26-30, 2003, pp. 26–30.

[31] H., Lilliefors, "On the Kolmogorov–Smirnov test for normality with mean and variance unknown", *Journal of the American Statistical Association*, vol. 62, pp. 399–402, 1967.

[32] F., Wilcoxon, "Individual comparisons by ranking methods", *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[33] H., Levene, "Robust tests for equality of variances" In Ingram Olkin; Harold Hotelling; et al. (eds.). *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford University Press. pp. 278–292, 1960.

[34] "Student" W. S. Gosset, "The probable error of a mean", *Biometrika*, vol. 6, vol. 1, pp. 1–25, 1908.

[35] B.L., Welch, "The generalization of 'Student's problem' when several different population variances are involved". Biometrika, vol. 34, no. 1–2, pp. 28–35, 1947.

Carlos Martín

Bachelor in Computer Science from UNED. He also holds an M.S. in Computer Science and Technology and Ph.D. in Computer Science from Universidad Carlos III de Madrid. His main interests in the field of Artificial Intelligence are focused on optimization techniques based on Evolutionary Computation. He is an engineer at the Security Operations Center of the Air Force JSTCIS Cyberdefense Directorate in The Spanish Ministry of Defense, Officer in charge of the Forensics, Intrusion Detection, Malware Analysis and Mitigation and Recovery sections.

David Quintana

Visiting Professor with the Department of Computer Science at Universidad Carlos III de Madrid, Spain. There, he is part the bio-inspired algorithms group EVANNAI. He holds a Bachelor in Business Administration and a Ph.D. in Finance from Universidad Pontificia Comillas (ICADE), a Bachelor in Computer Science from UNED and an M.S. in Intelligent Systems from Universidad Carlos III de Madrid. His current research interests are mainly focused on applications of evolutionary computation and artificial neural networks in finance and economics. David is former Chair of the Computational Finance and Economics Technical Committee of the IEEE Computational Intelligence Society.

Pedro Isasi

Graduate and Doctor in Computer Science by the Polytechnic University of Madrid since 1994. Currently he is University professor and head of the Evolutionary Computation and Neural Networks Laboratory in the Carlos III of Madrid University. Dr. Isasi has been Chair of the Computational Finance and Economics Technical Committee (CFETC) of the IEEE Computational Intelligence Society (CIS), Head of the Computer Science Department and Vice-chancellor in the Carlos III University among others. His research is centered in the field of the artificial intelligence, focusing on problems of Classification, Optimization and Machine Learning, fundamentally in Evolutionary Systems, Metaheuristics and Artificial Neural Networks.

# Attention-based Multi-modal Sentiment Analysis and Emotion Detection in Conversation using RNN

Mahesh G. Huddar[1,3]*, Sanjeev S. Sannakki[2,3], Vijay S. Rajpurohit[2,3]

[1] Department of Computer Science and Engineering, Hirasugar Institute of Technology, Nidasoshi, Belagavi (India)
[2] Department of Computer Science and Engineering, Gogte Institute of Technology, Belagavi (India)
[3] Visvesvaraya Technological University, Belagavi (India)

Received 6 June 2019 | Accepted 26 October 2020 | Published 1 December 2020



## Abstract

The availability of an enormous quantity of multimodal data and its widespread applications, automatic sentiment analysis and emotion classification in the conversation has become an interesting research topic among the research community. The interlocutor state, context state between the neighboring utterances and multimodal fusion play an important role in multimodal sentiment analysis and emotion detection in conversation. In this article, the recurrent neural network (RNN) based method is developed to capture the interlocutor state and contextual state between the utterances. The pair-wise attention mechanism is used to understand the relationship between the modalities and their importance before fusion. First, two-two combinations of modalities are fused at a time and finally, all the modalities are fused to form the trimodal representation feature vector. The experiments are conducted on three standard datasets such as IEMOCAP, CMU-MOSEI, and CMU-MOSI. The proposed model is evaluated using two metrics such as accuracy and F1-Score and the results demonstrate that the proposed model performs better than the standard baselines.

## Keywords

Attention Model, Interlocutor State, Contextual Information, Emotion Detection, Multimodal Fusion, Sentiment Analysis.

DOI: 10.9781/ijimai.2020.07.004

## I. Introduction

THE main aim of automatic sentiment and emotion analysis in conversational videos is to analyze and detect sentiment and emotional state of a participant in conversational videos. Due to the recent advancements in Internet technologies and social media networks, the users post their reviews, about a service or a product in the form of conversational videos on social media platforms, such as Twitter, Flicker, YouTube, and Facebook, etc. Recently, multimodal sentiment and emotion analysis from the conversation has become an interesting research topic due to its widespread applications in areas such as health-care assistant devices, education, dialogue understanding, human-computer interaction, and human resource management. In prior work, the unimodal features from the available modalities were extracted, and then the unimodal features are fused to form the multimodal feature vector. For multimodal fusion, there are three options, early fusion (feature concatenation), model-based fusion, or late (decision) fusion. In feature concatenation, the features from individual modalities are concatenated to get the multimodal feature vector.

Recently, many approaches were proposed for utterance level sentiment and emotion analysis [1], [2]. In late fusion, the feature vectors from individual modalities are modeled using the classical classifiers. The output of the classifiers on the unimodal feature vector is fused using an ensemble approach [3]. These fusion strategies perform fairly well but cannot accommodate the contextual information among the utterances and interlocutor state of the participant. More recently attention based contextual fusion and contextual cross modality fusion strategies show promising results. In the contextual fusion technique, the bidirectional recurrent neural network (RNN) was used to extract the context between the utterances of a video [4]. In contextual cross-modality fusion along with contextual information, the importance of modality is considered in multimodal fusion [5]. In [6] dynamic fusion is performed by paying attention at each time step. Evolutionary computing-based multi-layer feature optimization is used to improve the overall accuracy of classification in [7].

The sentiment or emotional state of the particular participant in the conversation is not considered for analysis in these models. Hence the existing models fail to capture the contextual information among the utterances and flow of conversation. But in reality, the contextual state and sentiment or emotion of a particular party does add a lot of value to the overall result. The proposed model believes that the sentiment or emotional state of an utterance mainly depends on the interlocutor state of the participant, the previous emotional state of the participant, and context between the utterances [8]. By incorporating the interlocutor state of the particular participant and context between the utterances, the results of the proposed method outperform the baselines by over 2%.

The main contributions of the proposed model are,

- An effective multimodal sentiment and emotion analysis technique is proposed to extract the contextual information among the utterances and accommodate the interlocutor state of a particular participant in the conversation.

* Corresponding author.
E-mail address: mailtomgh1@gmail.com

- The pair-wise attention-based mechanism is used to understand the relationship and importance of modalities before fusion.
- The proposed model effectively captures the sentiment or emotional state of the participant in the conversation.
- The model is tested and validated on three standard datasets and the results are compared against the standard baselines for multimodal sentiment and emotion analysis in conversational videos.

The structure of the remaining sections of the article is as follows: the important work carried-out in multimodal sentiment and emotion analysis, context extraction between the utterances and traditional techniques in multimodal fusion are described in Section II. The proposed attention-based multimodal sentiment and emotion analysis in the conversation using the RNN model is presented in Section III. The experimental setup, results on three standard datasets, and comparison of results against a standard baseline of the proposed model are presented in Section IV. Finally, future work in multimodal affective computing in conversational videos is presented and concludes the paper in Section V.

## II. Related Work

Sentiment analysis and Emotion detection in conversation are popular research topics in multimodal affective computing [9] because of their applications in various areas such as sentiment analysis, health-care assistance devices, recommendation systems, education, human-computer interaction, etc. [10]. The multimodal data has information in three modes such as text (transcribed audio), audio, and video. The traditional multimodal sentiment analysis and emotion detection technique extracts the unimodal features from the three modalities, use either feature level (early) fusion [11] [12] or decision level (late) fusion [13] [14] [15] or hybrid fusion [16] to merge effective information from different modalities.

An utterance is a segment or a part of the video (may not be a complete sentence) and video reviews contain a sequence of such multiple utterances. In utterance or segment level sentiment and emotion classification, each segment of a video is analyzed and assigned a label [17]. Recently, many approaches were proposed for analyzing sentiment and detecting emotion at the utterance level [1], [2]. In [18] authors extracted acoustic, lexicon, and visual features and used an ensemble approach to ensemble classification of SVM classifier. Their proposed ensemble approach achieves better results than conventional methods. Authors in [19] fused acoustic and linguistic cues at feature level using 3-D activation valance for emotion recognition. In [20] authors extracted textual, speech, and visual features using convolutional neural networks. They analyzed sentiment and emotion using multiple kernel learning.

In [21] acoustic information and visual cues are fused to model multimodal emotion recognition system and contextual information is used for sentiment and emotion analysis. In recent works on multimodal sentiment and emotion analysis in conversational videos, each utterance of a video is processed sequentially using RNN. The model proposed in [8] propagates the context among the utterances and sequential information to the next utterance. They use bidirectional recurrent neural networks [22] to extract the context between the utterances and feed the information sequentially. DialogueRNN [23] uses an attention-based pooling approach to capture the context of a particular utterance in the conversation. However, this pooling based attention mechanism fails to consider participant information of particular utterance and its effect on other utterances. They use a global state and participant state for modeling multimodal emotion detection in conversation.

Other notable works include [24] [25] [26] where multimodal sentiment and emotion detection is addressed using deep learning-based models. Ghosal et al. [27] proposed a pair-wise attention-based method to understand the importance of individual modalities and the relationship between the modalities before fusion. The two-dimensional graph-based feature extraction methods using fuzzy logic are discussed in [28] [29] and [30]. The PRAAT[1] software was used to extract the emotional state from voice [31]. The proposed model considers context between the utterances, the interlocutor state of a participant, and previous emotion state to effectively model the multimodal sentiment and emotion analysis system in conversational videos.

## III. Proposed Methodology

The proposed attention-based multimodal sentiment and emotion analysis in the conversation using RNN is discussed in detail in this section. The overview of the proposed model is:

- First, the utterance level features of individual modalities such as acoustic, textual, and visual features are extracted.
- The pair-wise attention-based mechanism is used to understand the relationship and importance of modalities before fusion.
- The gated recurrent unit (GRU), a variant of RNN, is used to model the interlocutor state of the participant, context extraction, and emotion decoding.
- Bimodal and trimodal fusion are performed by considering the previous emotional state, the importance of individual modality, and interlocutor state. A trimodal representation of feature vector acts as an input for final sentiment or emotion prediction.

### A. Dataset Used

The model is evaluated on three standard datasets, IEMOCAP [32], CMU-MOSEI [24] for multimodal emotion detection, and CMU-MOSI [33] for multimodal sentiment analysis.

#### 1. IEMOCAP

IEMOCAP dataset is a collection of 12-hours of two-way acted dyadic conversations among multiple speakers. The conversational video is divided into multiple opinion segments called utterances. Each of the utterances is annotated with emotion labels such as anger, sadness, excitement, happiness, fear, neutral, and surprise. Videos with angry, happy, sad, excited, frustrated, and neutral are considered to compare against the state of the art models.

#### 2. CMU-MOSEI

The CMU-MOSEI dataset contains 3228 videos with 23453 small segments called utterances from 1000 speakers collected from YouTube. CMU-MOSEI is a transcribed, gender-balanced, properly punctuated dataset. The average number of segments per video is 7.3 and the average length of each segment is 7.28 seconds. The total number of words and unique words in utterances are 447143 and 23026 respectively. The dataset is manually labeled with 6 emotions such as anger, disgust, fear, happiness, sadness, and surprise.

#### 3. CMU-MOSI

There are 93 videos with 2199 utterances in CMU-MOSI dataset where 89 speakers review various products and topics in English. The average length of a segment is 4.2 seconds and about 12 words per utterance. Each utterance is manually labeled by 5 assessors with a score ranging from -3 and +3. The average of these 5 assessors is taken as sentiment polarity. The Video/Utterance level Train-Test

[1] https://www.fon.hum.uva.nl/praat/

distributions of CMU-MOSEI, CMU-MOSI, and IEMOCAP datasets are shown in Table I. The Label distribution statistics of CMU-MOSI and IEMOCAP datasets are given in Table II and Table III respectively.

TABLE I. Video/Utterance Level Train-Test Distribution of CMU-MOSI, CMU-MOSEI, and IEMOCAP Dataset

| | Videos | | Utterances | |
|---|---|---|---|---|
| | Train | Test | Train | Test |
| CMU-MOSI | 62 | 31 | 1447 | 752 |
| CMU-MOSEI | 2583 | 646 | 18051 | 4625 |
| IEMOCAP | 120 | 31 | 5810 | 1623 |

TABLE II. Label Distribution Statistics of CMU-MOSI

| | Positive | Negative |
|---|---|---|
| CMU-MOSI | 1176 | 1023 |

TABLE III. Label Distribution Statistics of IEMOCAP

| | Neutral | Happy | Sadness | Anger | Frustrated | Excited |
|---|---|---|---|---|---|---|
| IEMOCAP | 1708 | 648 | 1084 | 1103 | 1849 | 1041 |

## B. Feature Extraction

This section discusses the steps followed in extracting features from acoustic, text, and visual modalities.

### 1. Audio Feature Extraction

OpenSMILE [34] open-source tool is used for acoustic feature extraction from CMU-MOSI, CMU-MOSEI, and IEMOCAP datasets. The acoustic features are extracted at a frame rate of 30Hz and 100ms sliding window. The dimension of utterance level features for acoustic modality is 73, 73, and 384 for CMU-MOSI, IEMOCAP, and CMU-MOSEI datasets respectively.

Let $f_{ai}$ be the feature vector of $i^{th}$ segment then, the acoustic feature vector $f_a$ is represented by,

$$f_a = < f_{a1}, f_{a2}, f_{a3}, \dots \dots, f_{an} > \quad (1)$$

Where n is the number of segments or utterances.

### 2. Textual Feature Extraction

Features from text (transcribed text) modality are extracted from each utterance using Convolutional Neural Networks (CNN) [35] from CMU-MOSI and IEMOCAP datasets. First, each utterance is represented as word2vec vectors [36] to understand the context in the text. These Word2Vec vectors are processed using 3 convolutional layers. The three layers have feature maps of size 50, 75, and 100 with filters of sizes 2, 3, and 2 respectively. Max-pooling of a 2x2 window size is used after every convolutional layer. The fully connected layer receives input from the convolution layer and output is fed to a softmax classifier. The fully connected layer has 600 neurons with ReLU [37] activation function. The softmax output of the convolutional neural network (CNN) is used as the textual features. GloVe embedding's used for extracting textual features from the CMU-MOSEI dataset. The dimension of utterance level features for textual modality is 100 for CMU-MOSI, IEMOCAP datasets, and 300 for CMU-MOSEI dataset.

Let $f_{ti}$ be the feature vector of $i^{th}$ segment then, the textual feature vector $f_t$ is represented by,

$$f_t = < f_{t1}, f_{t2}, f_{t3}, \dots \dots, f_{tn} > \quad (2)$$

Where n is the number of segments or utterances.

### 3. Visual Feature Extraction

In the past 3D convolutional neural networks have been successfully used for object detection and classification [38]. The results presented in [38], outperform the traditional object tracking and detection, and motivate us to adopt 3D-CNN in our work. Visual features are extracted using 3D-CNN from CMU-MOSI and IEMOCAP datasets and Facet[2] tool from the CMU-MOSEI dataset. The dimension of utterance level features for visual modality is 100 for CMU-MOSI, IEMOCAP datasets, and 35 for CMU-MOSEI dataset.

Let $f_{vi}$ be the feature vector of $i^{th}$ segment then, the visual feature vector $f_v$ is represented by,

$$f_v = < f_{v1}, f_{v2}, f_{v3}, \dots \dots, f_{vn} > \quad (3)$$

Where n is the number of segments or utterances.

## C. Problem Statement

Let $P_1$ and $P_2$ be the two participants in the conversation. The $u_1, u_2 \dots u_n$ are the utterances uttered by either of the participants $P_1$ and $P_2$ with sentiment score and one of the emotion labels such as happy, sad, anger, surprise, disgust, and fear is assigned to the utterances. As each of the utterances is uttered by either of the participants in the conversation, this allows capturing the average sentiment of the participant in sentiment score or emotion label calculation. Also, it avoids misclassification due to long pauses by the participant in the conversation. Let $u_t$ be the $t^{th}$ utterance uttered by the party $P_1$ or $P_2$ at timestamp t, which is represented by three modalities such as text, visual and acoustic,

$$u(p)_t = < t_t, v_t, a_t > \quad (4)$$

where $t_t$, $v_t$, and $a_t$ are textual, visual and acoustic feature vectors of the $t^{th}$ utterance at timestamp t and p Є P1, P2.

The objective function of the problem is to accept the feature vector from three modalities of an utterance, cumulative context representation of the conversation and emotional state of the previous participant, and output the sentiment score and associated emotion label.

## D. Proposed Model Description

The sentiment or emotion of an utterance depends on the cumulative contextual state of the conversation, the interlocutor state, and the sentiment or emotional state of the previous participant. Hence the proposed model considers the cumulative context and emotion of participants to predict the sentiment or emotional state of an utterance. The proposed model has three branches of recurrent neural networks (RNN) to capture the participant interlocutor state, cumulative context, and sentiment or emotional state of the participant. Each modality uses one RNN to capture participant dyadic information and another set of RNN's are used to capture the sentiment or emotional state of the participant. One RNN is used to capture the cumulative contextual information. A weighted-pooling based pairwise attention-based mechanism is performed to understand the relative importance of individual modalities before fusion. Finally, two-two modalities and then all modalities are fused to form a trimodal representation of feature vector for predicting the sentiment score or emotion label of an utterance.

### 1. Interlocutor State

The interlocutor state of the network captures and keeps track of the state of the participant involved in the multimodal conversation. The network has nxm number of RNN's, where n is the number of participants and m is the number of modalities. The output of the interlocutor state is the input for updating cumulative contextual vector and emotion or sentiment prediction of the utterance. Initially, the interlocutor state is initialized to the null vector. For the utterance at timestamp t the interlocutor state $i_t$ of a particular modality is updated $i_{t+1}$ using feature

[2] https://goo.gl/1rh1JN

representation of particular modality at timestamp t (that is $f(t)_t$ or $f(a)_t$ or $f(v)_t$) and attentive cumulative contextual vector representation until timestamp t (that is $C(t)_t$ or $C(a)_t$ or $C(v)_t$). The purpose of using the cumulative contextual vector along with utterance representation is to understand the contextual information of conversation until that timestamp. The steps in the interlocutor state update are described using the following formula and shown in Fig. 1.

$$i(m)_t = \text{Interlocutor}((f(m)_t \oplus C(m)_t), i(m)_{t-1}) \quad (5)$$

where $\oplus$ represents concatenation operator and m is the modality with values either t or a or v.

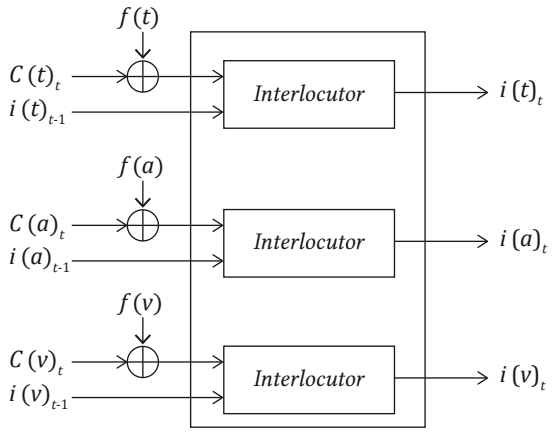



Fig. 1. Interlocutor State Update at timestamp t.

## 2. Cumulative Context

In conversational sentiment analysis and emotion detection, to determine the sentiment or emotional state of an utterance at timestamp t, the preceding utterances at time < t can be considered as its cumulative context. The interlocutor state of the previous utterance (that is $i(t)_{t-1}$ or $i(a)_{t-1}$ or $i(v)_{t-1}$) and utterance level modality representation at timestamp t (that is $f(t)_t$ or $f(a)_t$ or $f(v)_t$) are used to change the cumulative context vector representation from $c_{t-1}$ to $c_t$. This helps to understand the dependencies between the utterances and participants. The steps in the cumulative context state update are described using the following formula and shown in Fig. 2.

$$c(m)_t = \text{Context}((f(m)_t \oplus i(m)_{t-1}), c(m)_{t-1}) \quad (6)$$

where $\oplus$ represents concatenation operator and m is the modality with values either t or a or v.

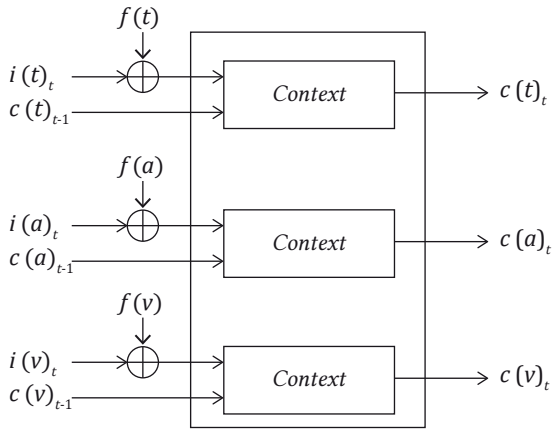



Fig. 2. Cumulative Context state update at timestamp t.

Weighted pooling based attention is performed over cumulative context vector representation until timestamp t.

$$C(m)_t = \frac{e^{c(m)_t}}{\sum_{k=1}^{t} e^{c(m)_k}} \quad (7)$$

Where, $C(m)_t$ is the attentive cumulative contextual vector.

## 3. Emotion State

The emotional state network is used to decode the sentiment or emotional information encoded by interlocutor state RRN. The previous emotion state output (that is $e(t)_{t-1}$ or $e(a)_{t-1}$ or $e(v)_{t-1}$) and interlocutor state sentiment or emotional information (that is $i(t)_t$ or $i(a)_t$ or $i(v)_t$) are the input to emotion state RNN at timestamp t. Weighted pooling based pair-wise attention is performed on the output produced by emotion state RNN to produce the relevant sentiment or emotion label. The steps in the emotion state update are described using the following formula and shown in Fig. 3.

$$e(m)_t = \text{emotion}(i(m)_t, e(m)_{t-1}) \quad (8)$$

where $\oplus$ represents concatenation operator and m is the modality with values either t or a or v.

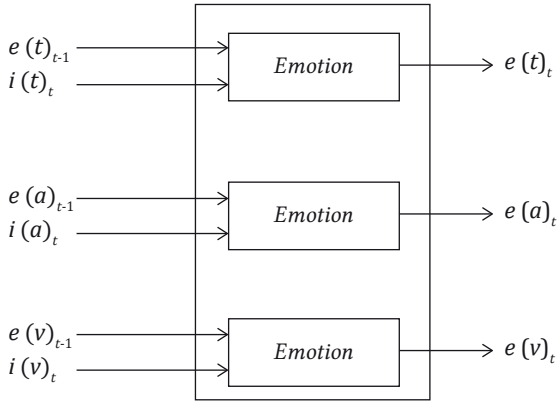



Fig. 3 Emotion State update at timestamp t.

## 4. Weighted Pooling Based Pair-Wise Attention and Bimodal Fusion

For each timestamp t, the emotion state network produces emotion vectors for each modality such as $e(t)_t$, $e(a)_t$ and $e(v)_t$. Weighted pooling based pair-wise attention [4] is performed between two-two emotion vectors at a time to get bimodal representation emotion vectors. Let X and Y be the two emotion state outputs produced by the emotion state network at timestamp t, then the weighted pooling based pair-wise attention mechanism is performed as follows:

$$M(t)_1 = X.Y^T \quad \text{and} \quad M(t)_2 = Y.X^T \quad (9)$$

$$W(t)_1 = \frac{e^{M(t)_1}}{\sum_{k=1}^{t} e^{M(k)_1}} \quad (10)$$

$$W(t)_2 = \frac{e^{M(t)_2}}{\sum_{k=1}^{t} e^{M(k)_2}} \quad (11)$$

$$O(t)_1 = W(t)_1 \,.\, Y \quad \text{and} \quad O(t)_2 = W(t)_2 \,.\, X \quad (12)$$

$$A(t)_1 = O(t)_1 \odot X \quad \text{and} \quad A(t)_2 = O(t)_2 \odot Y \quad (13)$$

$$B\_Fusion(XY)_t = A(t)_1 \oplus A(t)_2 \quad (14)$$

Where B_Fusion is the bimodal fusion at timestamp t.

The pair-wise matching matrices at timestamp t are calculated in equation (9), then the probability distribution scores (weights)

of each modality are calculated in equation (10) and (11). Modality specific attentive representations are calculated in equation (12). An important component among the multiple modalities and utterances is calculated by performing element-wise matrix multiplication as shown in equation (13). Attentive matrix representations are then concatenated to produce bimodal representation at timestamp t as shown in equation (14). The steps in Weighted Pooling based pair-wise attention and bimodal fusion at timestamp t are shown in Fig. 4.

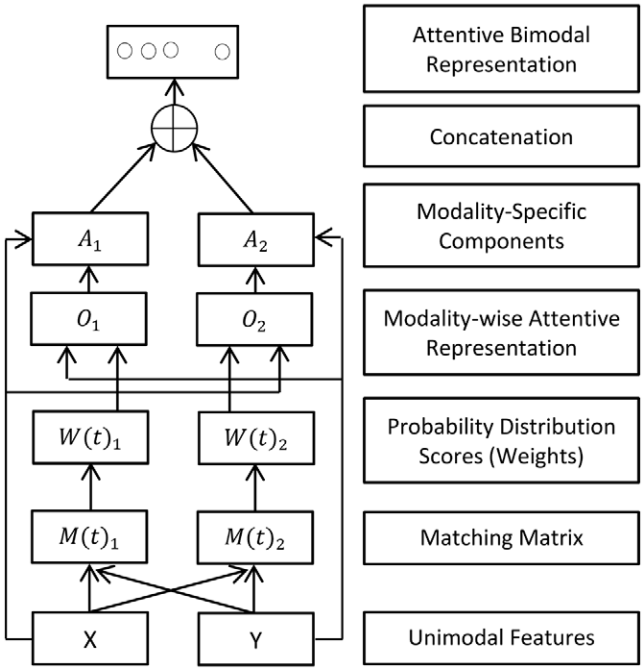



Fig. 4. Weighted Pooling based Pair-Wise Attention and Bimodal Fusion at timestamp t where X, Y Є {T, A, V}.

### 5. Trimodal Fusion

The bimodal attentive representation and emotional state of the utterance are used to get the trimodal representation. The bimodal attentive representation and output of emotion state RNN at timestamp t is concatenated to form the final trimodal attentive representation at timestamp t. The trimodal fusion at timestamp t is shown in Fig. 5.

$$e(tav)_t = e(t)_t \oplus e(a)_t \oplus e(v)_t \oplus B\_Fusion(TA) \oplus B\_Fusion(TV) \oplus B\_Fusion(VA) \quad (15)$$

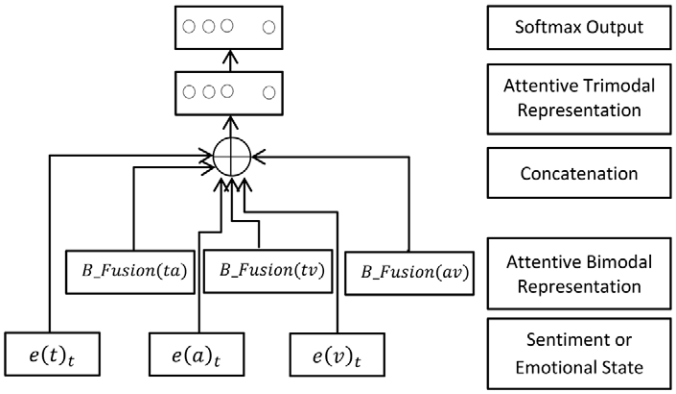



Fig. 5. Trimodal Fusion at timestamp t.

### E. Classification and Training

The trimodal sentiment or emotional representation is fed to the softmax classifier to predict the testing label $\hat{y}$ for an utterance in the conversation. The softmax classifier takes the concatenated sentiment or emotion vector $e(tav)_t$ at timestamp t as an input. The softmax output is represented as,

$$p(y|U) = \text{softmax}\left(w^{(s)}(e(tav)_t) + b^{(s)}\right) \quad (16)$$

Where $w^{(s)}$ is the weight matrix, $b^{(s)}$ is the bias matrix, p is a predicted sentiment or emotion class.

$$\hat{y} = \underset{y}{\text{argmax}}\ p(y|U) \quad (17)$$

Where $\hat{y}$, is the predicted label of testing utterance.

The cross-entropy loss function L(θ) is used to train the model and is represented as,

$$L(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_i^j \log\hat{y}_i^j + \lambda\sum_{k=0}^{N}\theta_k^2 \quad (18)$$

where $N$ is the number of utterances in training data. $y_s$ and $\hat{y}_s$ are the true and predicted label of the $s^{th}$ utterance. $M$ is the number of categories (classes) and λ is the L2-regularization term. Adam [39] is used to optimize the cross-entropy loss function parameters due to its ability to adapt to the learning rate for each learning parameter. The proposed algorithm for attention-based multimodal sentiment and emotion analysis in the conversation using RNN is summarized in Table IV.

## IV. Results Analysis and Discussion

The proposed attention-based multimodal sentiment and emotion analysis framework in the conversation using RNN is implemented in python using the PyTorch and tensor flow is used as backend. The model is evaluated on the Tesla K80 GPU with a 12GB RAM hardware configuration. The experiments are conducted on three standard datasets such as CMU-MOSI, CMU-MOSEI, and IEMOCAP. The experimental results of the proposed method are compared against the standard baselines such as [25] , [27], [40], [41], and [42]. The proposed model is evaluated using two metrics, classification accuracy, and F1-score. First, the results are obtained for the combination of two-two modalities such as text-audio, text-video, and audio-video, and then all three modalities with and without attention mechanism. The comparison of results of the proposed technique for sentiment analysis with and without attention is given in Tables V and VI. The results show that the attention-based model performs better than the standard baselines in all possible combinations of constituent modalities except for the audio-video combination on the CMU-MOSI dataset. The trimodal model performs better than the bimodal model. The emotion detection results on CMU-MOSEI and IEMOCAP datasets with and without attention are shown in Tables VII and VIII. The results show that the attention-based models are performing better than the standard baselines and model without attention except for the label happy in the CMU-MOSEI dataset. Fig. 6, Fig. 7, Fig. 8 and Fig. 9 show a comparison of the experimental results of the proposed method on CMU-MOSEI, IEMOCAP, CMU-MOSI datasets against standard baselines. On CMU-MOSI and CMU-MOSEI datasets the trimodal models are performing better than the bimodal and unimodal models, whereas A-V combination is performing the worst among all possible combination of models in sentiment classification. For emotion classification, the proposed model obtains the best results on the CMU-MOSEI dataset as it effectively uses all the available modalities and captures the contextual information since the availability of large dataset for training.

TABLE IV. ALGORITHM FOR PROPOSED MULTIMODAL SENTIMENT AND EMOTION CLASSIFICATION IN THE CONVERSATION USING RNN

| | |
|---|---|
| 1: *Procedure* $FeatureExtraction(U)$<br>2: *for* $i$ *in* 1 *to* $N$ *do*:<br>3: $f(t)_i \longleftarrow audioFeatures(U_i)$<br>4: $f(a)_i \longleftarrow textFeatures(U_i)$<br>5: $f(v)_i \longleftarrow videoFeatures(U_i)$ | *Procedure to extract unimodal features* |
| 6: *Procedure* $InterlocutorState(t, m)$<br>7: $i(m)_t = Intrerlocutor((f(m)_t \oplus C(m)_t), i(m)_{t-1})$<br>8: *return* $(i(m)_t)$ | *Procedure to update Interlocutor state at time t* $m \in \{t, a, v\}$ |
| 9: *Procedure* $ContextExtract(t, m)$<br>10: $c(m)_t = Context((f(m)_t \oplus i(m)_{t-1}), c(m)_{t-1})$<br>11: $\mathrm{C(m)_t} = \frac{e^{c(m)_t}}{\sum_{k=1}^{t} e^{c(m)_k}}$<br>12: $return(C(m)_t)$ | *Procedure to extract cumulative context* $m \in \{t, a, v\}$ |
| 13: *Procedure* $EmotionState(t, m)$<br>14: $e(m)_t = Emotion(i(m)_t, e(m)_{t-1})$<br>15: $return(e(m)_t)$ | *Procedure to update Emotion state at time t* $m \in \{t, a, v\}$ |
| 16: *Procedure* $Attention(t, X, Y)$<br>17: $M(t)_1 = X.\mathrm{Y^T}$ *and* $M(t)_2 = Y.\mathrm{X^T}$<br>18: $W(t)_1 = \frac{e^{M(t)_1}}{\sum_{k=1}^{t} e^{M(k)_1}}$<br>19: $W(t)_2 = \frac{e^{M(t)_2}}{\sum_{k=1}^{t} e^{M(k)_2}}$<br>20: $O(t)_1 = W(t)_1 . Y$ *and* $O(t)_2 = W(t)_2 . X$<br>21: $A(t)_1 = O(t)_1 \odot X$ *and* $A(t)_2 = O(t)_2 \odot Y$<br>22: $return(A(t)_1 \oplus A(t)_2)$ | *Procedure for weighted pooling based attention and fusion* |
| 23: *Procedure* $B\_Fusion(t, X, Y)$<br>24: $B(X, Y)_t = Attention(t, X, Y)$<br>25: $retrun(B(X, Y)_t)$ | *Procedure for Bimodal fusion at time t* |
| 26: *Procedure* $T\_Fusion(t)$<br>27: $e(tav)_t = e(t)_t \oplus e(a)_t \oplus e(v)_t \oplus B\_Fusion(TA) \oplus B\_Fusion(TV) \oplus B\_Fusion(VA)$<br>28: $retrun\ (e(tav)_t)$ | *Procedure for Trimodal fusion at time t* |
| 29: *Procedure* $Classification\ (U, t)$<br>30: *for* $i$ *in* 1 *to* $N$ *do*:<br>31: $p\ (y\|U) = softmax\ (w^{(s)}(e(tav)_t) + b^{(s)})$<br>32: $\hat{y} = \underset{y}{argmax}\ p\ (y\|U)$<br>33: $return\ (\hat{y})$ | *Procedure for classification of utterance into discrete number of classes* |
| 34: $FeatureExtraction(U)$ | *Unimodal Feature Extraction* |
| 35: *for* $X, Y \in \{t, a, v\}$<br>36: $B(X, Y)_t \longleftarrow B\_Fusion(t, X, Y)$ | *Bimodal Fusion* $X, Y \in \{t, a, v\}$ |
| 37: $e(tav)_t \longleftarrow T\_Fusion(t)$ | *Trimodal Fusion* |
| 38: $C_t \longleftarrow Classification\ (e(tav)_t)$ | *Classification* |

TABLE V. Experimental Results of the Proposed Method on CMU-MOSI Dataset Compared Against Standard Baselines, T for Text, A for Audio and V for Video

| Modality | Poria et al. [40] | Zadeh et al. [41] | GRU - Without Attention | | GRU - With Attention | |
|---|---|---|---|---|---|---|
| | Accuracy | Accuracy | Accuracy | F1-Score | Accuracy | F1-Score |
| T + A | 73.7 | 71.1 | 76.79 | 77.62 | 79.71 | 73.38 |
| T + V | 74.1 | 73.7 | 79.35 | 78.54 | 80.14 | 73.40 |
| A + V | 68.4 | 67.4 | 65.51 | 67.77 | 66.28 | 67.79 |
| T+A+V | 74.1 | 73.6 | 80.02 | 73.73 | 80.62 | 74.33 |

TABLE VI. Experimental Results of the Proposed Method on CMU-MOSEI Dataset Compared Against Standard Baselines

| Modality | Zadeh et al. [42] | Ghosal et al. [27] | GRU - Without Attention | | GRU - With Attention | |
|---|---|---|---|---|---|---|
| | Accuracy | Accuracy | Accuracy | F1-Score | Accuracy | F1-Score |
| T + A | - | 79.74 | 80.02 | 73.73 | 80.64 | 73.29 |
| T + V | - | 79.40 | 80.31 | 73.42 | 80.54 | 76.22 |
| A + V | - | 76.66 | 76.79 | 77.62 | 80.45 | 73.50 |
| T+A+V | 76.90 | 79.80 | 80.98 | 73.51 | 81.29 | 73.12 |

TABLE VII. Experimental Results of the Proposed Method on CMU-MOSEI Dataset Compared Against Standard Baselines with the T-A-V Combination of Modalities

| Label | Ghosal et al. [27] | | GRU - Without Attention | | GRU - With Attention | |
|---|---|---|---|---|---|---|
| | Accuracy | F1-Score | Accuracy | F1-Score | Accuracy | F1-Score |
| Anger | 62.60 | 72.80 | 81.49 | 74.20 | 83.58 | 76.10 |
| Fear | 62.00 | 89.90 | 88.92 | 84.83 | 91.20 | 87.01 |
| Happy | 66.30 | 66.30 | 58.51 | 44.30 | 59.79 | 44.74 |
| Sad | 60.40 | 66.90 | 76.93 | 67.86 | 78.90 | 69.60 |
| Surprise | 53.70 | 85.50 | 87.51 | 82.78 | 89.75 | 84.91 |
| Disgust | 69.10 | 76.60 | 88.92 | 84.83 | 91.20 | 87.01 |

TABLE VIII. Experimental Results of the Proposed Method on IEMOCAP Dataset Compared Against Standard Baselines with the T-A-V Combination of Modalities

| Label | Ghosal et al. [27] | | Hazarika et al. [25] | | GRU - Without Attention | | GRU - With Attention | |
|---|---|---|---|---|---|---|---|---|
| | Accuracy | F1-Score | Accuracy | F1-Score | Accuracy | F1-Score | Accuracy | F1-Score |
| Angry | 64.7 | 65.2 | 68.2 | 68.2 | 73.1 | 66.6 | 75.2 | 68.5 |
| Neutral | 58.5 | 59.2 | 59.9 | 60.6 | 79.8 | 76.1 | 82.1 | 78.3 |
| Happy | 25.6 | 33.1 | 23.6 | 32.8 | 52.3 | 39.1 | 53.9 | 40.7 |
| Sad | 75.1 | 78.8 | 70.6 | 74.4 | 69.2 | 61.1 | 71.0 | 62.6 |
| Excited | 80.2 | 71.8 | 72.2 | 68.4 | 78.3 | 74.1 | 80.8 | 76.4 |
| Frustrated | 61.1 | 58.9 | 71.9 | 66.2 | 79.6 | 75.9 | 82.1 | 78.3 |

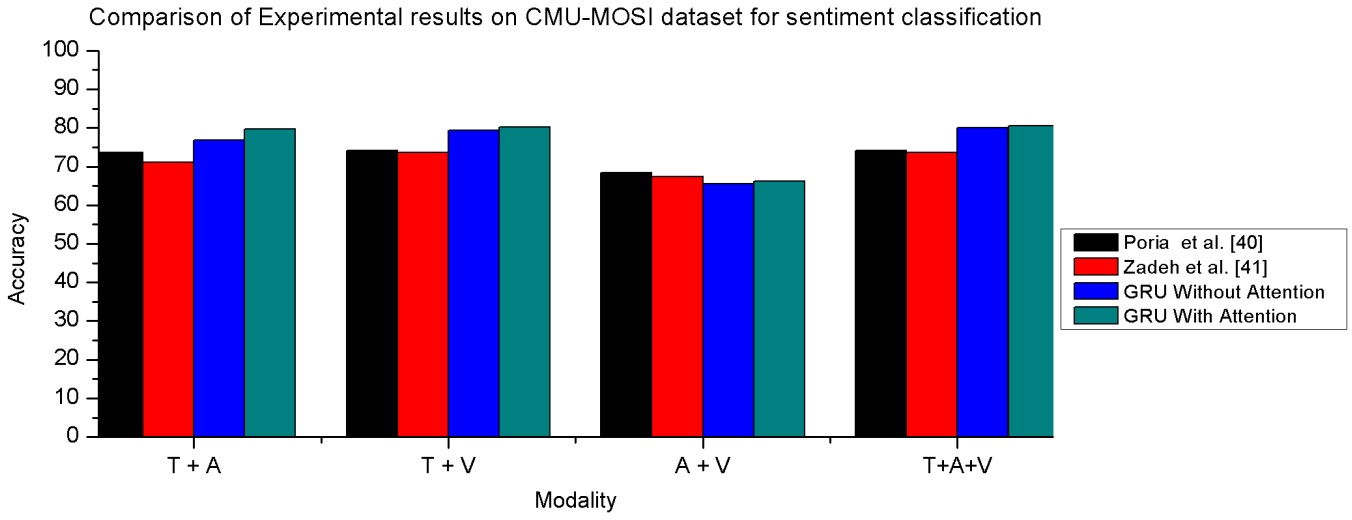



Fig. 6. Comparison of experimental results of the proposed method on CMU-MOSI dataset against standard baselines, Legend: T: Text, A: Audio, V: Video.

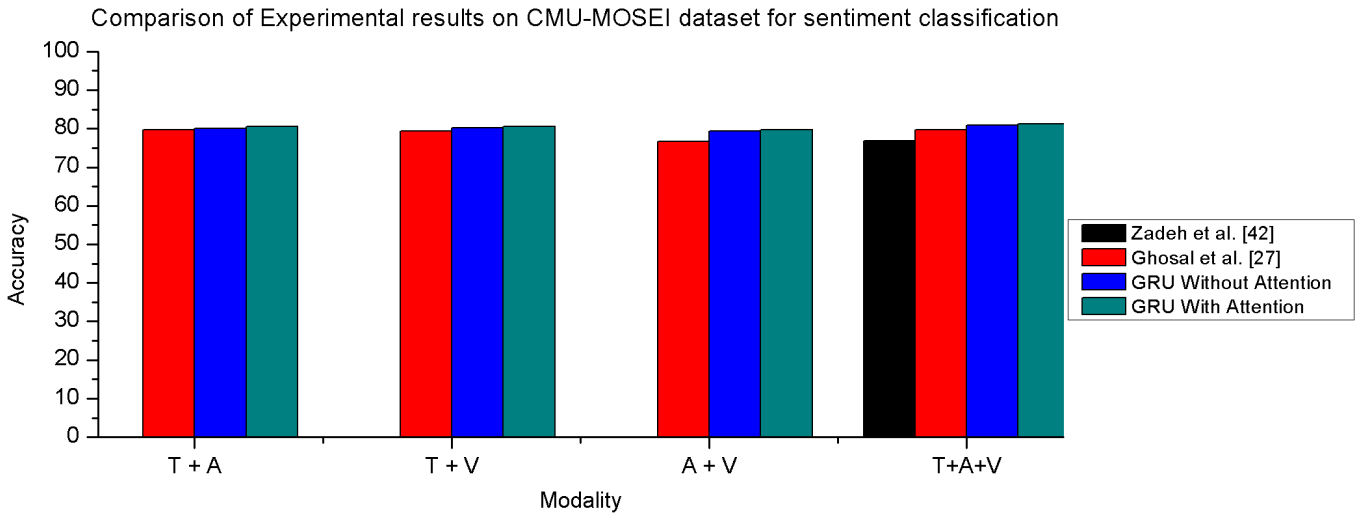



Fig. 7. Comparison of experimental results of the proposed method on CMU-MOSEI dataset against standard baselines.

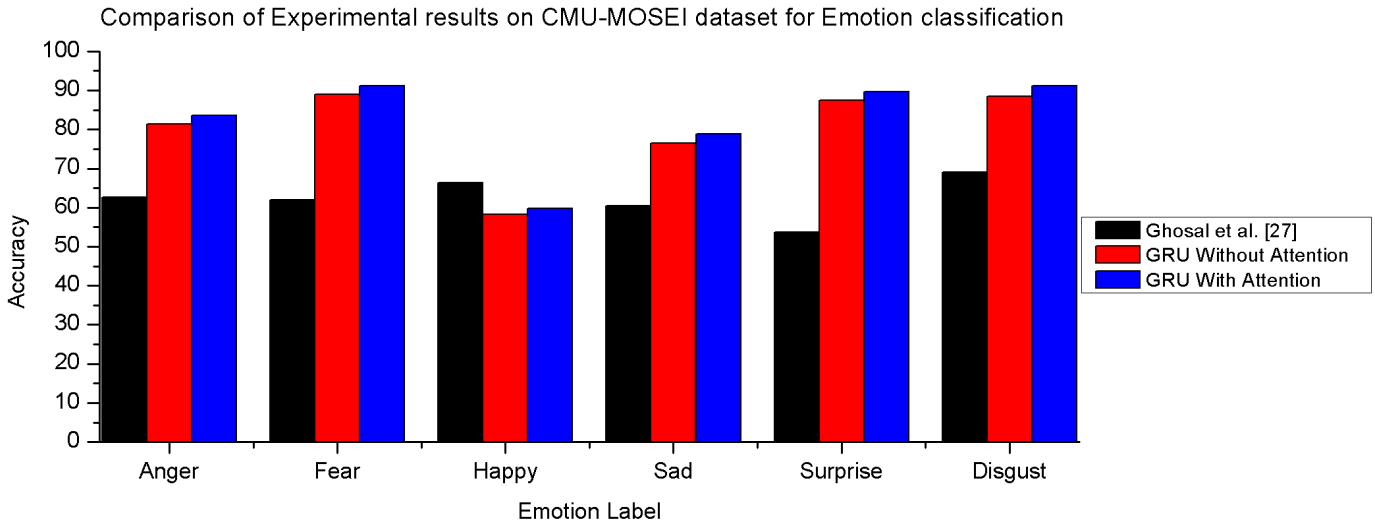



Fig. 8. Comparison of experimental results of the proposed method on CMU-MOSEI dataset against standard baselines.

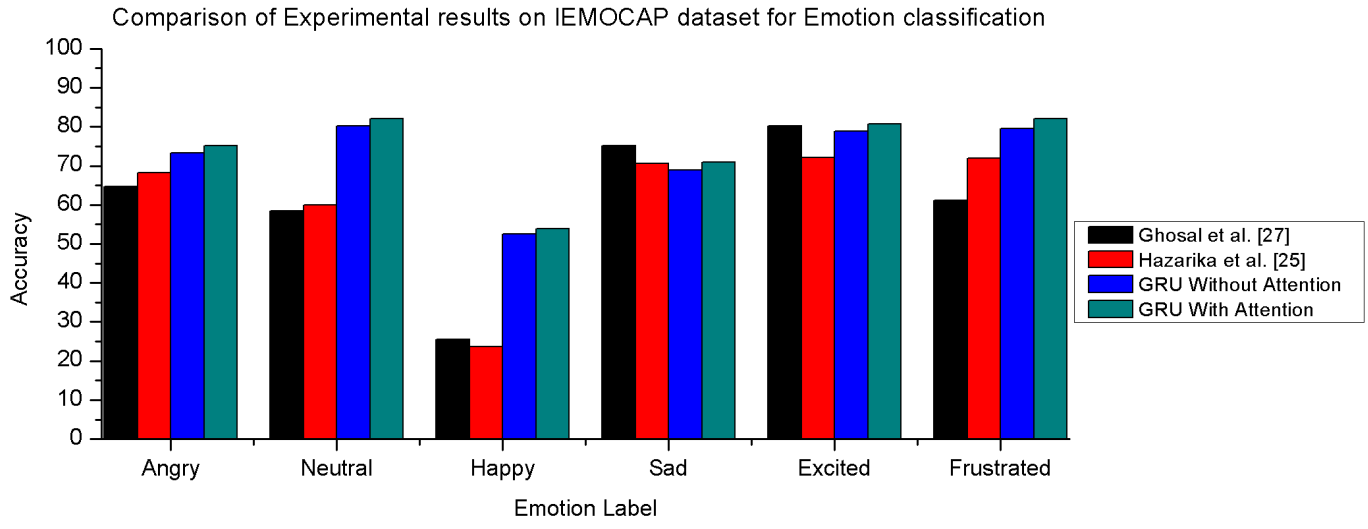



Fig. 9. Comparison of experimental results of the proposed method on IEMOCAP dataset against standard baselines

## V. Conclusion and Future Work

The multimodal fusion, capturing interlocutor state of the participant, and understanding context between the utterances are the most important issues in multimodal sentiment analysis and emotion detection in conversation. In this paper first, features from individual modalities such as textual, acoustic, and visual features are extracted. Textual features are extracted using CNN and GloVe embedding's, audio features using open smile toolkit and visual features using 3D-CNN and facet toolkit. An attention-based pair-wise technique is used to extract the context between the utterances and understand the importance of constituent modalities before fusion. The recurrent neural network, more specifically gated recurrent Unit (GRU) based model is used to capture the interlocutor state and context extraction. By incorporating contextual information, the interlocutor state, and previous emotion state, the proposed model performs better than the standard baselines in terms of classification accuracy. In the future, we will explore techniques to address more than two participants in conversational videos. Also, we will study the feature selection methods to understand whether the emotion-specific features can improve the overall classification accuracy.

## References


[1] S. Poria, E. Cambria and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2539–2544, 2015.

[2] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers), pp. 873-883, 2017.

[3] M. G. Huddar, S. S. Sannakki and V. S. Rajpurohit, "An Ensemble Approach to Utterance Level Multimodal Sentiment Analysis," in *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, Belgaum, India, pp. 145-150, 2018.

[4] M. G. Huddar, S. S. Sannakki and V. S. Rajpurohit, "Multi-level context extraction and attention-based contextual inter-modal fusion for multimodal sentiment analysis and emotion classification," International Journal of Multimedia Information Retrieval, vol. 9, no. 2, pp. 103-112, 2020, https://doi.org/10.1007/s13735-019-00185-8.

[5] M. G. Huddar, S. S. Sannakki and V. S. Rajpurohit, "Attention-based word-level contextual feature extraction and cross-modality fusion for sentiment analysis and emotion classification," *International Journal of Intelligent Engineering Informatics*, vol. 8, no. 1, pp. 1-18, 2020.

[6] S. Chen and Q. Jin, "Multi-modal conditional attention fusion for dimensional emotion prediction," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016.

[7] M. G. Huddar, S. S. Sannakki and V. S. Rajpurohit, "Multi-level feature optimization and multimodal contextual fusion for sentiment analysis and emotion classification," *Computational Intelligence*, vol. 36, no. 2, pp. 861-881, 2020.

[8] S. Poria, N. Majumder, R. Mihalcea and E. Hovy, "Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances," *arXiv preprint arXiv:1905.02947*, 2019.

[9] B. Kratzwald, S. Ilić, M. Kraus, S. Feuerriegel and H. Prendinger, "Deep learning for affective computing: Text-based emotion recognition in decision support," *Decision Support Systems*, vol. 115, pp. 24-35, 2018.

[10] M. G. Huddar, S. S. Sannakki and V. S. Rajpurohit, "A Survey of Computational Approaches and Challenges in Multimodal Sentiment Analysis," *International Journal of Computer Sciences and Engineering*, vol. 7, no. 1, pp. 876-883, 2019.

[11] V. P. Rosas, R. Mihalcea and L.-P. Morency, "Multimodal sentiment analysis of Spanish online," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 38 - 45, 2013.

[12] V. Perez-Rosas, R. Mihalcea and L.-P. Morency, "Utterance-Level Multimodal Sentiment Analysis," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, 2013.

[13] J. G. Ellis, B. Jou and S.-F. Chang, "Why We Watch the News: A Dataset for Exploring Sentiment in Broadcast Video News," in *Proceedings of the 16th International Conference on Multimodal Interaction*, Istanbul, Turkey, 2014.

[14] F. Celli, B. Lepri, J.-I. Biel, D. Gatica-Perez, G. Riccardi and F. Pianesi, "The Workshop on Computational Personality Recognition 2014," in *Proceedings of the 22nd ACM international conference on Multimedia*, Orlando, Florida, USA , 2014.

[15] H. Kumar and B. Harish, "Automatic Irony Detection using Feature Fusion and Ensemble Classifier," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 7, pp. 70-79, 2019.

[16] H. Kumar, B. Harish and H. Darshan, "Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 5, pp. 109-114, 2019.

[17] M. G. Huddar, S. S. Sannakki and V. S. Rajpurohit, "Multimodal Emotion Recognition using Facial Expressions, Body Gestures, Speech, and Text Modalities," *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 8, no. 5, pp. 2453-2459, 2019.

[18] V. Rozgić, S. Ananthakrishnan, S. Saleem, R. Kumar and R. Prasad, "Ensemble of SVM trees for multimodal emotion recognition," in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, Hollywood, CA, USA, 2013.

[19] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie and R. Cowie, "On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues," *Journal on Multimodal User Interfaces*, vol. 3, no. 1-2, p. 7–19, 2010.

[20] S. Poria, I. Chaturvedi, E. Cambria and A. Hussain, "Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis," in *IEEE 16th International Conference on Data Mining (ICDM)*, Barcelona, Spain, 2016.

[21] D. Datcu and L. J. Rothkrantz, "Semantic audio-visual data fusion for automatic emotion recognition," *Emotion recognition: a pattern analysis approach*, pp. 411-435, 2014.

[22] J. Chung, C. Gulcehre, K. Cho and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," *arXiv:1412.3555*, 2014.

[23] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh and E. Cambria, "DialogueRNN: An Attentive RNN for Emotion Detection in Conversations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[24] A. Zadeh, P. P. Liang, S. Poria, E. Cambria and L.-P. Morency, "Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph," in *Proceedings of the 56th Annual Meeting of the As-sociation for Computational Linguistics (Volume 1: Long Papers)*, 2018.

[25] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria and R. Zimmermann, "ICON: interactive conversational memory network for multimodal emotion detection," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[26] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, Louis-Philippe Morency and R. Zimmermann, "Conversational Memory Network for Emotion Recognition in Dyadic Dialogue Videos," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long Papers), New Orleans, Louisiana, 2018.

[27] D. Ghosal, M. S. Akhtar, D. Chauhan, S. Poria, A. Ekbal and P. Bhattacharyya, "Contextual inter-modal attention for multi-modal sentiment analysis," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[28] M. Wan, G. Yang, S. Gai and Z. Yang, "Two-dimensional discriminant locality preserving projections (2DDLPP) and its application to feature extraction via fuzzy set," *Multimedia tools and applications*, vol. 76, no. 1, pp. 355-371, 2017.

[29] M. Wan, M. Li, G. Yang, S. Gai and Z. Jin, "Feature extraction using two-dimensional maximum embedding difference," *Information Sciences*, vol. 274, pp. 55-69, 2014.

[30] M. Wan, Z. Lai, G. Yang, Z. Yang, F. Zhang and H. Zheng, "Local graph embedding based on maximum margin criterion via fuzzy set," *Fuzzy Sets and Systems*, vol. 318, pp. 120-131, 2017.

[31] M. Magdin, T. Sulka, J. Tomanová and M. Vozár, "Voice Analysis Using PRAAT Software and Classification of User Emotional State," *IJIMAI*, vol. 5, no. 6, pp. 33-42, 2019.

[32] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335-359, 2008.

[33] A. Zadeh, R. Zellers, E. Pincus and L.-P. Morency, "Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages," *Journal IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82-88, 2016.

[34] F. Eyben, M. Wöllmer and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, Barcelona, Spain, 2013.

[35] A. Karpathy, G. Toderici, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-scale Video Classification with Convolutional Neural Networks," in *Proceedings of International Computer Vision and Pattern Recognition*, 2014.

[36] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *arXiv:1301.3781*, 2013.

[37] Y. W. Teh and G. E. Hinton, "Rate-coded restricted Boltzmann machines for face recognition," in *Proceedings of the 13th International Conference on Neural Information Processing Systems*, Cambridge, MA, USA, 2000.

[38] S. Ji, W. Xu, M. Yang and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221 - 231, 2013.

[39] D. a. B. J. Kingma, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*, vol. 15, 2014.

[40] S. Poria, I. Chaturvedi, E. Cambria and A. Hussain, "Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis," in *2016 IEEE 16th international conference on data mining (ICDM)*, Barcelona, Spain, 2016.

[41] A. Zadeh, M. Chen, S. Poria, E. Cambria and L.-P. Morency, "Tensor Fusion Network for Multimodal Sentiment Analysis," in *Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 2017.

[42] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), Melbourne, Australia, 2018.

Mahesh G Huddar

Mahesh G. Huddar is working as an Assistant Professor in Computer Science and Engineering at Hirasugar Institute of Technology, Nidasoshi, Belagavi, India and he is currently pursuing Ph.D. studies at the Visvesvaraya Technological University, Belagavi, India in the Department of Computer Science and Engineering. He received his Master and Bachelor of Science degrees from the Visvesvaraya Technological University, Belagavi, India in 2014 and 2008, respectively. He has published a many papers in journals, International, and National conferences. His main research interests include Machine Learning, Deep Learning, Multimodal Sentiment Analysis, and Multimodal Emotion Detection. He is a member of the IEEE.

Sanjeev S Sannakki

Prof. Sanjeev S. Sannakki is working as a Professor in the Department of Computer Science and Engineering at Gogte Institute of Technology, Belgaum, Karnataka, India. He pursued his B.E. in Electronics & Communication from Karnataka University Dharwad in 2004, M.Tech, and Ph.D. from VTU, Belagavi in 2009, and 2016 respectively. His research areas include Image Processing, Cloud Computing, Computer Networks, and Data Analytics. He has published a good number of papers in journals, International, and National conferences. He is the reviewer for a few international journals and conferences. He is also a life member of CSI, and ISTE associations.

Vijay S Rajpurohit

Prof. Vijay S Rajpurohit is working as a Professor in the Department of Computer Science and Engineering at Gogte Institute of Technology, Belgaum, Karnataka, India. He pursued his B.E. in Computer Science and Engineering from Karnataka University Dharwad, M.Tech from N.I.T.K Surathkal, and Ph.D. from Manipal University, Manipal in 2009. His research areas include Image Processing, Cloud Computing, and Data Analytics. He has published a good number of papers in journals, International, and National conferences. He is the reviewer for a few international journals and conferences. He is the associate editor for two international journals and a Senior Member of the International Association of CS and IT. He is also the life member of SSI, ISC, and ISTE associations.

# A Word Embedding Based Approach for Focused Web Crawling Using the Recurrent Neural Network

P. R. Joe Dhanith[1]*, B. Surendiran[1], S. P. Raja[2]

[1] Department of CSE, National Institute of Technology Puducherry, Karaikal (India)
[2] Department of CSE, Vel Tech Rangarajan Dr.Sagunthala R & D Institute of Science and Technology (India)

Received 11 March 2020 | Accepted 4 July 2020 | Published 25 September 2020

## Abstract

Learning-based focused crawlers download relevant uniform resource locators (URLs) from the web for a specific topic. Several studies have used the term frequency-inverse document frequency (TF-IDF) weighted cosine vector as an input feature vector for learning algorithms. TF-IDF-based crawlers calculate the relevance of a web page only if a topic word co-occurs on the said page, failing which it is considered irrelevant. Similarity is not considered even if a synonym of a term co-occurs on a web page. To resolve this challenge, this paper proposes a new methodology that integrates the Adagrad-optimized Skip Gram Negative Sampling (A-SGNS)-based word embedding and the Recurrent Neural Network (RNN).The cosine similarity is calculated from the word embedding matrix to form a feature vector that is given as an input to the RNN to predict the relevance of the website. The performance of the proposed method is evaluated using the harvest rate (hr) and irrelevance ratio (ir). The proposed methodology outperforms existing methodologies with an average harvest rate of 0.42 and irrelevance ratio of 0.58.

## Keywords

Focused Crawler, Semantic Similarity, Word Embedding, Adagrad, Cosine, Recurrent Neural Network.

DOI: 10.9781/ijimai.2020.09.003

## I. Introduction

There has been a rapid increase in the number of web pages, from only 20 million in 2010 to 1.7 billion in 2020 [1]. The exponential increase in the volume of web pages each year has made it difficult for search engines to index them [2]–[5]. At the heart of a search engine is a web crawler, which is a software bot that retrieves web pages, commencing from seed URLs. A classic web crawler retrieves huge masses of information from the internet for a search, including information on irrelevant topics. Classic web crawlers demand huge storage capacities as well as additional downloading time. Such a problem calls for a topic-driven focused crawler that only downloads relevant web pages from the internet for a given topic.

Fig. 1 illustrates the working design of a focused web crawler, wherein the initial URLs are set by users for a given topic. The crawler visits web pages from initially-defined URLs and computes the similarity score of unexplored web pages. Based on the relevance score, precedence is assigned and stored in a web page archive.

Most focused web crawlers [6]–[9] only use full-page text to compute the similarity score of a web page, while others [10]–[13] use both full-page and anchor texts to calculate the relevance score, and [14]–[16] use cosine similarity to calculate the similarity score of unvisited web pages. In numerous existing studies [14]–[16], the cosine similarity value is calculated by finding the TF-IDF. The TF-IDF-based cosine similarity calculates the relevance score only if the topic term co-occurs with the terms on the web page, or else the similarity value is set to zero. The cosine similarity-based focused crawler provides a zero similarity score if the web page is semantically related but with no terms in common between the web page and the topic.

It was to overcome such challenges that researchers began working on ontology learning-based crawlers [17]–[19] to establish the semantic similarity between the topic and web page. Domain-specific ontologies are designed by domain experts, and crawlers fetch wrong results when human errors or discrepancies occur in the ontologies.

This paper proposes a new word embedding-based approach using the RNN to resolve this issue. Word embedding is one of the most common web page vocabulary representations. It is capable of capturing the meaning of a word on a web page, its semantic and syntactic similitude, and its relationship with other words. This work uses the A-SGNS to handle rare words that show up in the vocabulary. From the embedding matrix, the cosine similarity between the topic and web pages is calculated. The calculated cosine vectors are given as input to the RNN to predict the relevance of the web page.

The major contributions of this paper are as follows:

(1) Integrating the A-SGNS model with the RNN to automatically retrieve relevant web pages,

(2) Optimizing the SGNS using the Adagrad algorithm and the RNN using the RMSprop algorithm, and

(3) Implementing and evaluating the following nine different focused crawlers, including the breadth-first search (BFS)-based, vector space model (VSM), ontology learning-based using artificial neural network (ANN), Naive Bayes (NB)-based, link context-based, ANN-

* Corresponding author.
E-mail address: joe.dhanith@gmail.com

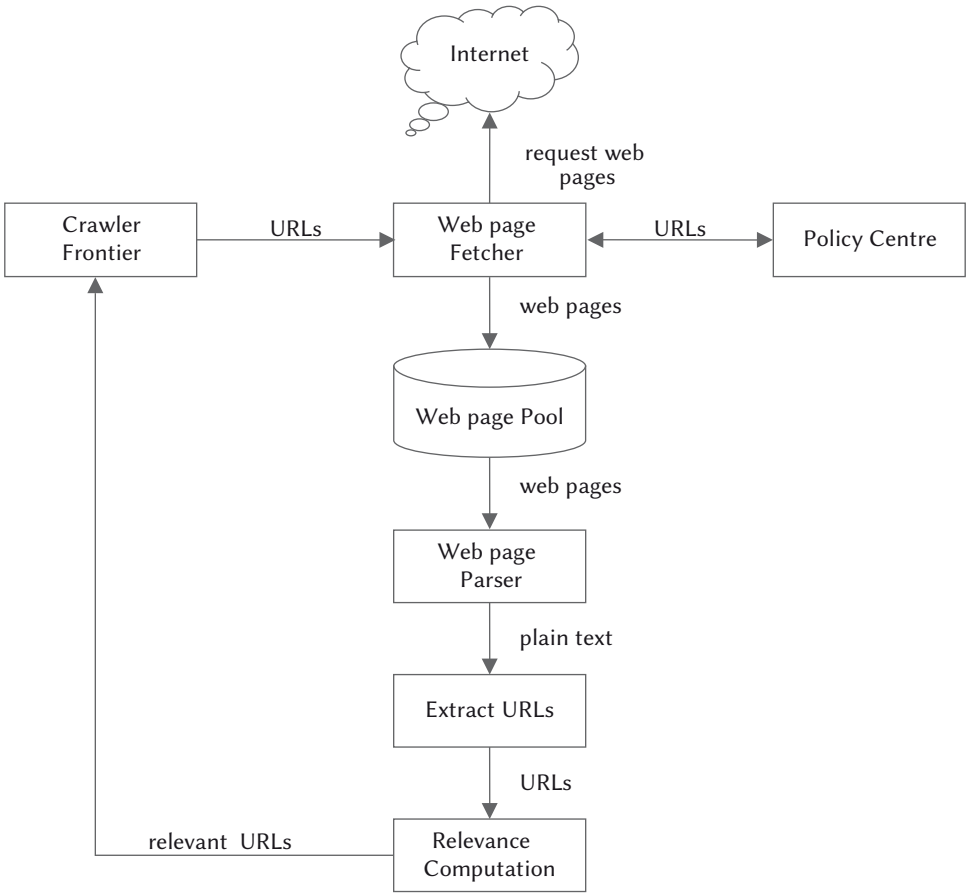



Fig. 1. Working design of focused web crawler.

based, semi-supervised, optimized Naive Bayes (ONB)-based, and the proposed RNN crawler. The efficiency of the nine focused crawlers is assessed using the harvest rate and irrelevance ratio.

The remainder of this article is organized as follows: Section II addresses existing methodologies. Section III describes the newly-constructed crawler framework, and Section IV the experimental design. Section V presents the experimental analysis, and Section VI concludes the paper.

### *A. Working of Focused Web Crawler*

Fig. 1 shows the working architecture of the focused web crawler, whose major components are a crawler frontier, web page fetcher, policy centre, web page pool, web page parser and relevance computation. The crawler frontier is a priority queue that stores a list of URLs in a prioritized order, based on the relevance score. The web page fetcher downloads web pages. The policy centre checks whether the web page is downloadable, and the web page pool stores downloaded web pages. The web page parser parses the web page to plaintext. The relevance computation computes the relevance of unvisited web pages. The stepwise working of the focused web crawler is as follows:

**Step 1**:The crawler frontier is initialized with the seed URLs and the policy centre with the depth of the web pages explored.

**Step 2**: The web page fetcher downloads web pages in the crawler frontier one by one. Once a web page is downloaded, the web page fetcher extracts the URLs present therein and sends them to the policy centre.

**Step 3**: The policy centre checks the downloadability of the received URL. A downloadable URL is sent back to the web page fetcher for downloading, else it is terminated. Steps (2) and (3) are repeated until the user-defined depth is reached.

**Step 4**: Once the web page fetcher receives the URL from the policy centre, it downloads the web page and stores it in the web page pool as a HTML document.

**Step 5**: The stored web pages are sent to the web page parser, along with the URLs, to retrieve meaningful information.

**Step 6**: The extracted information snippets are despatched to the relevance computation module to determine the relevance of the web page to the given topic. If the web page is relevant, the extracted URL is sent to the crawler frontier or else it is terminated.

## II. Related Work

The Google search engine uses the PageRank algorithm [2] to compute the relevance score of web pages. The PageRank algorithm is a voting method based on the number of incoming links and the rank of incoming links. If the number and the rank of the incoming links are high, the PageRank of the web page is also correspondingly high. The Google PageRank algorithm is formulated as follows in Equation (1),

$$p(w_p) = (1 - d_f) + d_f \cdot \left(\frac{p(l_1)}{c(l_1)} + \frac{p(l_2)}{c(l_2)} + \cdots + \frac{p(l_n)}{c(l_n)}\right) \quad (1)$$

where $d_f$ is the damping factor, the value of $d_f$ is usually set to 0.85, $p(w_p)$ is the PageRank of the web page $(w_p)$, $l_1, l_2, \ldots, l_n$ are the incoming links to the web page $w_p$, $p(l_1)$ is the PageRank of the first incoming link $(l_1)$, $c(l_1)$ is the number of outgoing links from web page $l_1$.

The baseline focused crawler downloads only relevant web pages by computing the relevance score of target variables such as full-page terms and anchor terms. The priority of the unvisited hyperlinks is calculated by combining the relevance score of the target variables. The priority score is given as a cosine function, as shown in Equation (2),

$$f_p(url) = \frac{f_{rs}(t, p) + f_{rs}(t, a)}{2} \quad (2)$$

where $f_p(url)$ is the priority function of unvisited URL, $f_{rs}(t, p)$ is the cosine function between the given topic and the full page terms, $f_{rs}(t, a)$ is the cosine function between the given topic and the anchor terms.

Andrea Capuano et al. [20] designed an ontology learning-based focused crawler using the convolution neural network (CNN). This work uses the Dbpedia spotlight and ImageNet to annotate, respectively, web page text and image data. A Li [21] semantic similarity algorithm calculates the textual relevance between the topic and the web page, and a CNN algorithm computes the relevance score between the downloaded image and the image in the knowledge base. The classification of the text and image is combined to identify the relevance of the web page. This work produced an average harvest rate of 0.29 after 5000 web page downloads.

Javad Hosseinkhani et al. [22] proposed an ontology learning-based focused crawler using the ant colony optimization (ACO) algorithm. The crime ontology builder in this work is used to design a crime ontology repository that annotates web pages. The ACO crawls and prioritizes web pages from the internet, while the support vector machine (SVM) classifies the relevance of a particular web page.

Debajyoti Mukhopadhyay et al. [23] advanced a semantic focused crawler to download relevant URLs. This work proposed a relevance score calculation formula between the topic and the web page, as shown in Equation (3),

$$f_{rs} = \sum f_o . N_o + \sum f_s . N_s \quad (3)$$

where $f_{rs}$ is the relevance score function, $f_o$ is the ontology-based term weight, $N_o$ is the count of the terms in the ontology, $f_s$ is the synonym weight value of the term, and $N_s$ is the count of the synonyms in the ontology. A web page with a relevance score above 0.5 is considered relevant, otherwise it is not.

Juan Qiu et al. [24] designed a focused crawler for the OpenStack Questions and Answers (Q&A) knowledge base. This work uses the linear discriminant analysis (LDA) clustering algorithm to construct the QA knowledge base topic corpus. A VSM is applied to find the similarity between the topic and the web page for corpus update.

Tanaphol SUEBCHUA et al. [25] propounded a history feature-based focused crawler. The history feature is extracted to reduce the priority score of the unvisited web page that downloads irrelevant web pages consecutively. The history feature, along with the relevance score of the link context and page text, is given as input to a NB classifier to predict the relevance of the web page.

Guangxia Xu et al. [26] elucidated a focused crawler based on particle swarm optimization (PSO). Initially, the TF-IDF is applied to calculate the weight of the terms, following which the PSO is applied to predict the relevance of the web page.

H.Dong et al. [27] discussed a self-adaptive semantic focused (SASF) crawler that combines an information content (IC)-based semantic similarity measure and a statistics-based similarity measure to determine the similarity score of a web page with respect to a given topic. The relevance score is calculated only for a full-page text feature with the given topic.

The relevance score of the SASF [27] is computed using Equation (4).

$$f_{rs}(url) = \max\left(f_{ic}(t, p), f_{stsm}(t, p)\right) \quad (4)$$

where $f_{rs}(url)$ relevance score function of the URL, $f_{ic}(t, p)$ is the Information content based semantic similarity function between the given topic and the full page terms, $f_{stsm}(t, p)$ is the statistical based string matching between the given term and the full page terms.

Ya Jun et al. [28] designed a cell-like membrane computing optimization (CMCFC) algorithm. The relevance score is calculated for four target variables (web page contents, link context, title term and the surrounding paragraph text) with the given term, using the cosine-based similarity metric. The relevance score of the four target variables is combined to compute the precedence of the unexplored web pages.

The relevance score of the CMCFC [28] is computed using Equation (5),

$$f_{rs}(url) = f_{rs}(t, p) + f_{rs}(t, a) + f_{rs}(t, title) + f_{rs}(t, st) \quad (5)$$

where $f_{rs}(url)$ is the relevance score function of the URL, $f_{rs}(t, p)$ is the cosine function between the given topic and the full page terms, $f_{rs}(t, a)$ is the cosine function between the given topic and the anchor terms, $f_{rs}(t, title)$ is the cosine function between the given topic and the terms, rs (t, st) is the cosine function between the given topic and the surrounding terms.

Hai-Tao Zheng et al. [19] elucidated a semantic focused crawler based on the artificial neural network (ANN). The relevance score is calculated, based on the distance between the full-page text and the given topic in the ontology. The crawler computes the term frequency of the unvisited web pages and feeds them as input to the ANN. The relevance score of the unvisited web pages is the output of the ANN. A major limitation of this approach is its inability to work well in an uncontrolled web environment.

Ahmed I Saleh et al. [17] designed a focused crawler using the optimized NB classifier (ONB). This work integrates the NB classifier with the SVM to form an optimized NB classifier. An integrated SVM and genetic algorithm is used to remove outliers in the training samples. The training data with the outliers removed is thereafter used to train the NB classifier. The ONB finds the sense of the data using the $D^2O$ ontology, calculates the similarity score of the web page, and determines its relevance.

Hai Dong et al. [18] designed a semi-supervised ontology learning-based approach for focused web crawling. This work extracts the Resnik semantic similarity score [29] and the statistical-based co-occurrence similarity score between the topic and the web page contents as features. The feature vector is then given as input to the SVM classifier to predict the relevance of the web page.

A review of the literature revealed the following drawbacks:

1. The TF-IDF weighting scheme finds the relevance of a web page only if the topic term co-occurs in target variables such as web page text and anchor text. The relevance score is otherwise calculated as zero, given that the TF-IDF does not consider the semantic similarity of the web page.
2. If the number of words on a web page is high, the dimension of the feature space generated by the TF-IDF is also high. The dimensionality of the feature space using the TF-IDF depends on the number of words on the web page. The TF-IDF vectors of web pages cause the high-dimensionality feature space that results in inaccurately-performing NB, SVM and ANN classifiers in a crawling environment.
3. Learning ontological concepts during a dynamic crawling process is an expensive, time-consuming process for basic learning algorithms like the NB, SVM and ANN. The complexity of learning ontological concepts in a crawling environment culminates in existing ontology learning-based crawlers performing at levels below par.

To solve these issues, this paper proposes a new word embedding-based approach using the RNN. An A-SGNS model is used to build a

VSM for representing words through a low-dimensional space. From the derived matrix, the cosine similarity between the extracted topics and the extracted web page terms is calculated to form a feature vector. The generated cosine feature vector is given as an input to the RNN to predict the relevance of the web page.

## III. Proposed Methodology

Fig. 2 shows the workflow diagram of the proposed work, with a six-layer architecture. The crawler frontier is initialized with the seed URLs and the topic by the user. The first layer is the topic preprocessing layer, where the given topic is preprocessed by applying methodologies like tokenization, Parts-of-Speech (POS) tagging, nonsense word filtering, stemming and synonym searches. The preprocessing is done using the Python Natural Language Toolkit (NLTK) library [30], [31], and the synonyms of the given topic are extracted for a meaningful search. The preprocessed topic terms are stored in a storage. The second layer is the crawling layer, where web pages are downloaded from the web, starting from manually assigned seed URLs. Once the download is done, the web pages are sent to the term extraction layer. The third layer is the term extraction layer, where the web pages are parsed to plaintext by removing HTML tags. After the parsing, target variables such as web page text and anchor text are extracted from the web page. The fourth layer is the term preprocessing layer, where the extracted target variables are preprocessed by applying methodologies like tokenization, POS tagging, nonsense word filtering and stemming. The preprocessing is carried out using the NLTK library [30], [31]. The fifth layer is the feature extraction layer, where the A-SGNS-based word embedding matrix is formed. From the derived matrix, the cosine similarity between the extracted topics and the extracted terms is calculated to form a feature vector. The generated cosine feature vector is given as input to the classification layer, where the recurrent neural network classifies the web page to determine its relevance.

### A. Recurrent Neural Networks

H. Palangi et al., [32]–[34] proposed a RNN for sentence embedding. The RNN, a type of deep learning model, uses the previous output as input in the hidden state and maintains the previous output to predict the current output. Fig. 3 shows the RNN workflow architecture of the proposed work.

The hidden state can be formulated as shown in equation (6):

$$s(t) = \tanh(w_{in}x(t) + w_{rec}s(t-1) + bias_s) \quad (6)$$

where $w_{in}$ is the input weight vector, $w_{rec}$ is the recurrent weight vector, s(t) is the hidden state, x(t) is the input vector, s(t–1) is the previous hidden state and $bias_s$ is the bias.

The output can be formulated as in equation (7):

$$o(t) = \sigma(w_{out}s(t) + bias_{out}) \quad (7)$$

where $\sigma(.)$ is the sigmoid activation function, o(t) is the output vector, $w_{out}$ is the output weight, and $bias_{out}$ is the bias of the output state.

### B. Feature Extraction

The first step in this work is to build a VSM to represent words through a low-dimensional space, using prediction-based word embedding. The A-SGNS [35]–[37], a prediction-based model which follows the neural network approach, is used. Given a sample of vocabulary, V, and the retrieved word context pair set, Z, let p(Z=1|(w, c)) be the likelihood that (w, c) arrives from Z and let p(Z=0|(w, c)) be the likelihood that (w, c) may not. The presupposition of SGNS is that if c is the context of word w in a window, the conditional probability of p(Z=1|(w, c))should be high, and otherwise small. Let $v_w$ denote the vector representation of w, and $v_c$ denote the vector of c. $D$ is the set of all pairs (w, c) that are in the text and $D'$ is the set of all pairs (w, c) not in the text. Then, $Z$ can be represented as follows in equation (8).

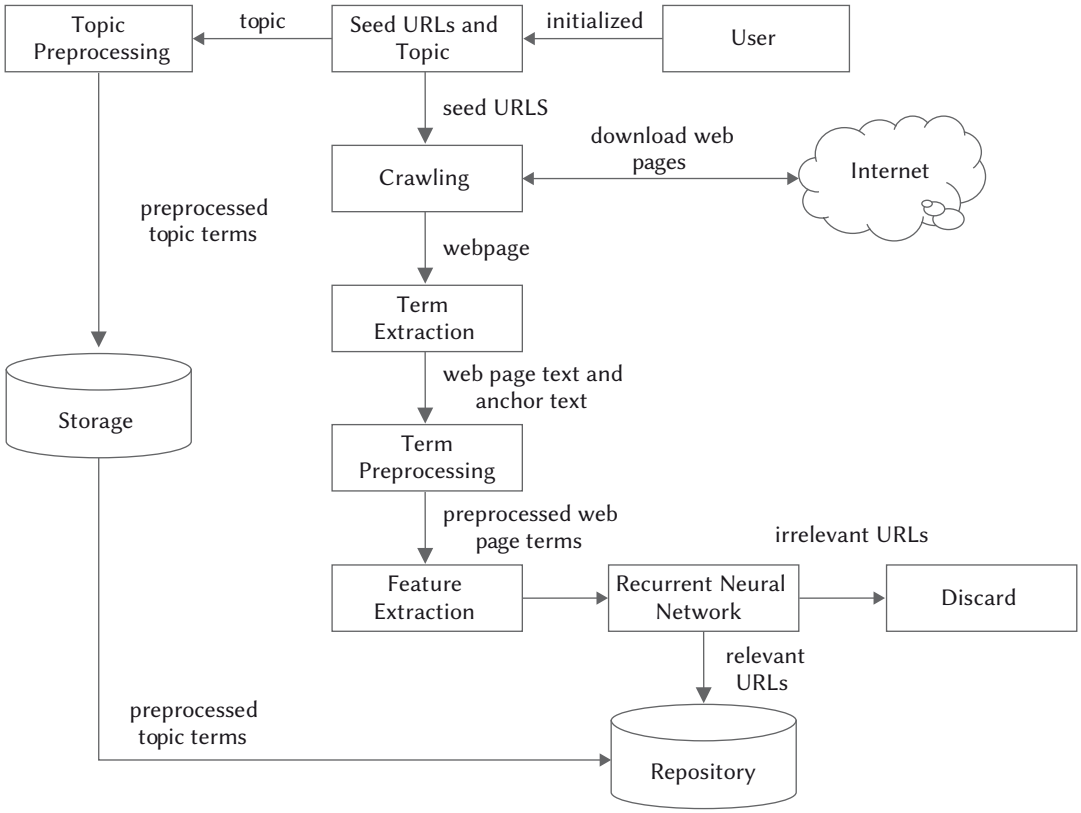



Fig. 2. Proposed workflow architecture.

$$Z = \begin{cases} 1 & if\ (w, c)\ \in D \\ 0 & if\ (w, c)\ \in D' \end{cases} \tag{8}$$

Then the p(Z=1|(w, c)) and p(Z=0|(w, c)) are computed as shown in equations (9) and (12) respectively,

$$p(Z = 1|(w, c)) = \frac{1}{1 + e^{v_c{}^T . v_w}} \tag{9}$$

$$p(Z = 0|(w, c)) = 1 - \frac{1}{1 + e^{v_c{}^T . v_w}} = \frac{e^{v_c{}^T . v_w}}{1 + e^{v_c{}^T . v_w}} \tag{10}$$

the above equation (10) is multiplied using $\frac{e^{-v_c{}^T . v_w}}{e^{-v_c{}^T . v_w}}$ and the following equations (11) and (12) are formulated.

$$p(Z = 0|(w, c)) = \frac{e^{v_c{}^T . v_w}}{1 + e^{v_c{}^T . v_w}} . \frac{e^{-v_c{}^T . v_w}}{e^{-v_c{}^T . v_w}} \tag{11}$$

$$p(Z = 0|(w, c)) = \frac{1}{1 + e^{-v_c{}^T . v_w}} \tag{12}$$

$$p(Z|(w, c); \theta) = \left(\frac{1}{1 + e^{v_c{}^T . v_w}}\right)^Z . \left(\frac{1}{1 + e^{-v_c{}^T . v_w}}\right)^{1-Z} \tag{13}$$

where $\theta=(v_c, v_w)$

$$L(\theta) = \prod_{(v_w, v_c) \in D \cup D'} \left(\frac{1}{1 + e^{v_c{}^T . v_w}}\right)^Z . \left(\frac{1}{1 + e^{-v_c{}^T . v_w}}\right)^{1-Z} \tag{14}$$

log likelihood is applied on both sides in equation (14) and formulated in the following equation (15).

$$l(\theta) = \sum_{(v_w, v_c) \in D \cup D'} Z . \log\left(\frac{1}{1 + e^{v_c{}^T . v_w}}\right) + (1 - Z) . \log\left(\frac{1}{1 + e^{-v_c{}^T . v_w}}\right) \tag{15}$$

$$l(\theta) = \sum_{(v_w, v_c) \in D} \log\left(\frac{1}{1 + e^{v_c{}^T . v_w}}\right) + \sum_{(v_w, v_c) \in D'} \left(\log\left(\frac{1}{1 + e^{-v_c{}^T . v_w}}\right)\right) \tag{16}$$

Let $\sigma(v_c{}^T . v_w) = \frac{1}{1+e^{v_c{}^T . v_w}}$ and $\sigma(-v_c{}^T . v_w) = \frac{1}{1+e^{-v_c{}^T . v_w}}$ then the equation (16) can be formulated as follows in equation (17),

$$l(\theta) = \log\left(\sigma(v_c{}^T . v_w)\right) + \sum_{i=1}^{k} E_{v_w \sim p_n(w)} \left[\log\left(\sigma(-v_c{}^T . v_w)\right)\right] \tag{17}$$

where $p_n(w) = \frac{\#w}{\sum_{w'} \#w'}$

Then the cost function can be given as follows in equation (18):

$$J(\theta; v_w, v_c) = \sum_{v_w \in w} \sum_{v_c \in w} \#(v_w, v_c) . \log\left(\sigma(v_c{}^T . v_w)\right) + \sum_{i=1}^{k} E_{v_w \sim p_n(w)} \left[\log\left(\sigma(-v_c{}^T . v_w)\right)\right] \tag{18}$$

The goal of any machine learning model is to find the optimal values of a weight matrix (θ) to minimize prediction errors. To update the lower learning rates for frequent words and higher learning rates for infrequent words, this work uses the Adagrad algorithm [38]–[40] for optimizing the cost function. The gradient descent on the cost function is applied with respect to $\theta$, as shown in Equation (19):

$$\frac{\partial J}{\partial \theta} = \#(v_w, v_c) . (\sigma(v_c{}^T . v_w) - 1) . \frac{\partial v_c{}^T . v_w}{\partial \theta} + \sum_{i=1}^{k} \sigma(-v_c{}^T . v_w) . \frac{\partial v_c{}^T . v_w}{\partial \theta} \tag{19}$$

A general AdaGrad update equation for cost function can be given in the following equation (20):

$$\theta = \theta - \frac{\omega}{\sqrt{\sum_{\tau=1}^{t-1} \frac{\partial J}{\partial \theta}^2}} \frac{\partial J}{\partial \theta} \tag{20}$$

From the designed word embedding model, the cosine similarity between the given topic and the web page is extracted as a feature. The cosine similarity between the topic and the content of the web page is given as follows in Equation (21):

$$sim(t, d) = \frac{\vec{t}^T . \vec{d}}{\|\vec{t}\| . \|\vec{d}\|} \tag{21}$$

where $\vec{t}^T$ is the embedding vector corresponding to the Topic, $\vec{d}$ is the embedding vector corresponding to the web page contents.

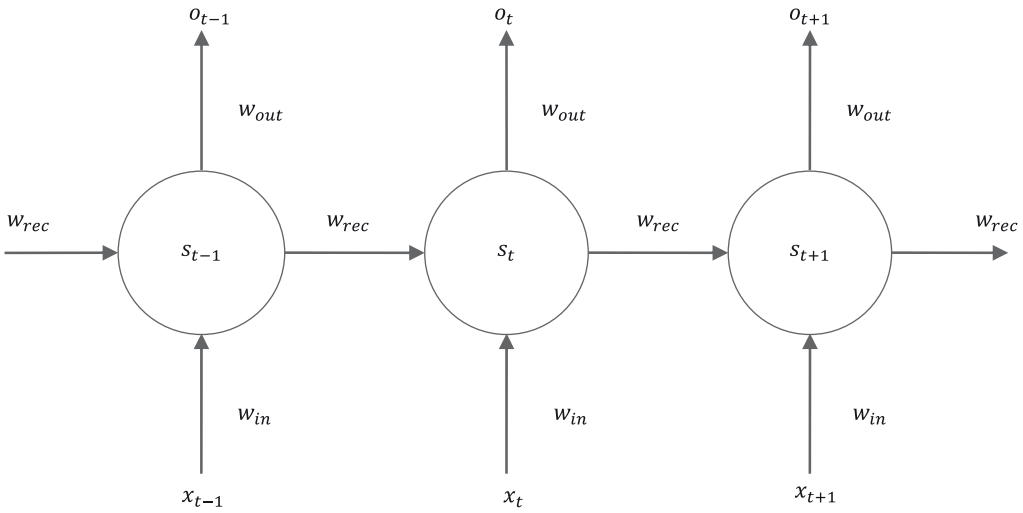



Fig. 3. Recurrent Neural Network workflow architecture of the proposed work.

### C. Classification Layer

The embedding vector of a document $\vec{d}$ is represented as $\{\overrightarrow{d_p}, \overrightarrow{d_{n_1}}, \overrightarrow{d_{n_2}}, \ldots, \overrightarrow{d_{n_m}}\}$.

where $\overrightarrow{d_p}$ is the positive sample among the web page documents and $\overrightarrow{d_{n_k}}$ is the kth negative sample of the same. These semantic vectors are produced by feeding the web page documents into the neural network (RNN), as discussed in section IIIA.

#### 1. Recurrent Weight

To maximize the likelihood of the positive document for the given document with respect to recurrent weight ($w_{rec}$) can be formulated as follows in equation (22):

$$L(w_{rec}) = \min_{w_{rec}} \left\{ -\log \prod_{i=1}^{N} P(\overrightarrow{d_p}|\vec{t}) \right\} \tag{22}$$

where $w_{rec}$ is the recurrent weight, $P(\overrightarrow{d_p}|\vec{t})$ is the probability of positive web page document for the ith Topic, and N is the number of topic-document pair in the corpus.

The above equation can be rewritten as follows in equation (23):

$$L(w_{rec}) = \min_{w_{rec}} \sum_{i=1}^{n} l_i(w_{rec}) \tag{23}$$

The $l_i(w_{rec})$ can be determined using the formula below from (24)-(28):

$$l_i(w_{rec}) = -\log\left(\frac{e^{\gamma.sim(t_i,d_i^+)}}{e^{\gamma.sim(t_i,d_i^+)} + \sum_{j=1}^{n} e^{\gamma.sim(t_i,d_{i,j}^-)}}\right) \tag{24}$$

$$l_i(w_{rec}) = \log\left(\frac{e^{\gamma.sim(t_i,d_i^+)} + \sum_{j=1}^{n} e^{\gamma.sim(t_i,d_{i,j}^-)}}{e^{\gamma.sim(t_i,d_i^+)}}\right) \tag{25}$$

$$l_i(w_{rec}) = \log\left(1 + \sum_{j=1}^{n} e^{\gamma.sim(t_i,d_{i,j}^-)} . e^{-\gamma.sim(t_i,d_i^+)}\right) \tag{26}$$

$$l_i(w_{rec}) = \log\left(1 + \sum_{j=1}^{n} e^{-\gamma.\left(sim(t_i,d_i^+) - sim(t_i,d_{i,j}^-)\right)}\right) \tag{27}$$

$$l_i(w_{rec}) = \log\left(1 + \sum_{j=1}^{n} e^{-\gamma.\Delta_{i,j}}\right) \tag{28}$$

where $\Delta_{i,j} = sim(t_i, d_i^+) - sim(t_i, d_i^-)$, the $\Delta_{i,j}$ value lies between 0 to 1, and γ is a scaling factor to increase the range of $\Delta_{i,j}$.

To perform back propagation through time [41] for $L(w_{rec})$ with respect to recurrent weight ($w_{rec}$) can be derived as follows in equation (29):

$$\frac{\partial L(w_{rec})}{\partial w_{rec}} = \sum_{i=1}^{n} \frac{\partial l_i(w_{rec})}{\partial w_{rec}} \tag{29}$$

The derived cost value of recurrent weight ($w_{rec}$) can be given as follows in equation (30):

$$\frac{\partial L(w_{rec})}{\partial w_{rec}} = \sum_{i=1}^{N} \sum_{j=1}^{n} \sum_{t=1}^{T} \alpha_{i,j,t} . \frac{\partial \Delta_{i,j,t}}{\partial w_{rec}} \tag{30}$$

where T is the number of time steps that the network is unfold over time and

$$\alpha_{i,j,t} = \frac{-\gamma . \sum_{j=1}^{n} e^{-\gamma.\Delta_{i,j,t}}}{1 + \sum_{j=1}^{n} e^{-\gamma.\Delta_{i,j,t}}}$$

The recurrent weight can be updated by using the RMSprop algorithm [42] because of its ability to update the lower learning rates for frequent parameters and higher learning rates for infrequent parameters and also clip the gradient when it goes higher than a threshold.

$$E[g^2]_t = \alpha E[g^2]_{t-1} + (1-\alpha)\left(\frac{\partial L(w_{rec})}{\partial w_{rec}}\right)^2 \tag{31}$$

$$w_{rec} = w_{rec} - \frac{\omega}{\sqrt{E[g^2]_t}} \frac{\partial L(w_{rec})}{\partial w_{rec}} \tag{32}$$

where $E[g^2]$ is the mean square of the gradient, α is the moving average parameter which is usually set to 0.9, ω is the learning rate which is set to 0.001, at each time step τ for the parameter $w_{rec}$.

#### 2. Input Weight

As derived for recurrent weight, the cost value of input weight ($w_{in}$) can be derived as follows in equation (33):

$$\frac{\partial L(w_{in})}{\partial w_{in}} = \sum_{i=1}^{N} \sum_{j=1}^{n} \sum_{t=1}^{T} \alpha_{i,j,t} . \frac{\partial \Delta_{i,j,t}}{\partial w_{in}} \tag{33}$$

## IV. Experimental Design and Analysis

A prototype of the crawlers (BFS, VSM, SVM, NB, ANN, ontology learning-based using the ANN, semi-supervised using the SVM, ONB-based and, finally, the proposed RNN) was developed in Python3 [43], [44], within the Spyder3.6 [45] platform. A cluster of six systems, each with the following configurations, was used to implement the prototypes: (i) 2.20GHz Intel Core i7-8750H 8th Gen processor, (ii) 16GB DDR4 RAM, (iii) 1TB serialATA hard drive, (iv) NVidia GeForce GTX 1060 6GB graphics, and (v) the Windows 10 operating system. These prototypes were implemented to crawl from the real web, using the Python packages, BeautifulSoup [46] and urllib [47]. BeautifulSoup package was used to handle HTML documents and urllib package was used to handle the URLs. The lxml parser [48] of BeautifulSoup package was used to parse the HTML documents. The urllib.parse function was used to parse the URLs. A set of ten topics and their respective seed URLs, as shown in Table 1, were given as input to all the crawlers. We collected 350000 (175000 positive and 175000 negative samples) URLs, along with their web page contents, for the topics shown in Table I in order to train the machine learning algorithms.

The experimental evaluations were carried out in two stages. The first stage was the training-testing phase of the machine learning algorithms, where the NB+TF-IDF, SVM+TF-IDF, ANN + TF-IDF and the proposed RNN + A-SGNS crawlers were evaluated using the metrics in Section V(A). The second stage was the crawling phase, where the performance of the crawlers (BFS, VSM, ontology learning-based using the ANN, NB-based, link context-based using the SVM, ANN-based, semi-supervised using the SVM, optimized Naive Bayes-based and the proposed RNN+A-SGNS) was evaluated using the metrics in Section V(C).

The NB, SVM and ANN algorithms, along with the TF-IDF, were implemented using the sci-kit learn Python package [49]. The NB-based crawler was implemented using the Gaussian Naive Bayes (GNB) classifier with a Laplace smoothing function, and the SVM-based crawler using a degree 1 linear SVM. The ANN model with 4 hidden nodes was implemented using the stochastic gradient descent (SGD) optimizer with the initialized weight value of 0.5 and learning rate of 0.1. In the proposed RNN model, the recurrent weight ($w_{rec}$) was initialized to -1.5 and the input weight ($w_{in}$) was initialized to 2.0. The learning rate ω was initialized to 0.001 for both $w_{rec}$ and $w_{in}$.

TABLE I. Seed URLs for the Ten Topics

| S.No | Topic | Seed URL |
|---|---|---|
| 1 | Football | • https://en.wikipedia.org/wiki/Football<br>• https://www.bbc.co.uk/sport/football |
| 2 | Knowledge Mapping | • https://www.apqc.org/blog/4-step-guide-knowledge-mapping<br>• https://www.mindmeister.com/blog/build-knowledge-map/ |
| 3 | Robot Army | • https://en.wikipedia.org/wiki/Military_robot<br>• https://www.popularmechanics.com/technology/robots/a29610393/robot-soldier-boston-dynamics/ |
| 4 | Smart Phone | • https://en.wikipedia.org/wiki/Smartphone<br>• https://www.amazon.in/Smartphones/b?ie=UTF8&node=1805560031 |
| 5 | Cloud Computing | • https://en.wikipedia.org/wiki/Cloud_computing<br>• https://azure.microsoft.com/en-in/overview/what-is-cloud-computing/ |
| 6 | wildfires | • https://en.wikipedia.org/wiki/Wildfire<br>• https://simple.wikipedia.org/wiki/Wildfire |
| 7 | Shahrukh khan | • https://en.wikipedia.org/wiki/Shah_Rukh_Khan<br>• https://www.imdb.com/name/nm0451321/ |
| 8 | computer | • https://en.wikipedia.org/wiki/Computer<br>• https://www.webopedia.com/TERM/C/computer.html |
| 9 | Apple | • https://www.apple.com/in/<br>• https://minecraft.gamepedia.com/Apple |
| 10 | Movie | • https://www.amctheatres.com/movies<br>• https://www.imdb.com/chart/moviemeter/ |

## V. Performance Evaluation

### A. Performance Evaluation of Training Phase

#### 1. Performance Metrics

This work uses four different metrics to measure the efficiency, at the training phase of different machine learning algorithms. They are accuracy (a), precision (p), recall (r) and F1-score (f) as shown in the following Equations (34), (35), (36), and (37) respectively.

$$a = \frac{tp + tn}{tp + tn + fp + fn} \tag{34}$$

$$p = \frac{tp}{tp + fp} \tag{35}$$

$$r = \frac{tp}{tp + fn} \tag{36}$$

$$f = \frac{2 * p * r}{p + r} \tag{37}$$

where *tp*, *tn*, *fp* and *fn* are true positive, true negative, false positive and false negative respectively.

### B. Analysis of Training Phase

A series of experiments was conducted to identify the right classifier with the requisite ability to guide the focused crawler. A dataset with 350,000 positive query-document pairs was collected for 10 different topics, as shown in Table I, each with 17,500 positive and 17,500 negative samples. Initially we applied tokenization, POS tagging, nonsense word filtering and stemming on both query and document data. The preprocessing was carried out using the Python Natural Language Toolkit (NLTK) [30], [31]. The nltk.word_tokenize() function was used to tokenize the topic words and the document words, the nltk.pos_tag() function to find the part of speech of each topic word and document word, and the nltk.stem package to find the root word of each topic word and document word. The words identified without POS tag were removed as non-sense words. Following the preprocessing of the training data, the TF-IDF-based cosine similarity and A-SGNS-based cosine similarity were extracted as a feature for each query-document pair. The TF-IDF-based extracted feature was used to train the NB, SVM, and ANN classifiers, while the A-SGNS-based extracted feature was used to train the RNN classifier. After training the classifiers, a testing dataset of 2827 query-document pairs was used to test the performance of the classifiers. The training phase was evaluated using four well-known metrics, formulated in Equations (34)-(37). Table II shows the results of a comparison of the four classifiers with 350,000 training data samples. The SVM with the TF-IDF, NB with the TF-IDF, ANN with the TF-IDF, and RNN with the A-SGNS produced accuracy of 0.623, 0.62, 0.70 and 0.813, respectively.

Logistic regression works well with linear data but not so with non-linear data. To predict categorical outcomes, it needs each data point to be independent. Given the limitations involved, it was, consequently, unable to perform well on the dynamic internet. Since the number of words in the web page was high, the dimensions created by the TF-IDF vectors were also high. In a high-dimensional feature space, the NB, SVM and ANN were affected by problems with overfitting and time consumption [50]. The NB, SVM and ANN failed to handle high-dimensional feature vectors and produced inaccurate results. The RNN, on the other hand, is a discriminative model that tries to differentiate between positive and negative samples in order to undertake the classification. In the proposed work, the A-SGNS model was used to build a VSM to represent words through a low-dimensional space. The ability of the RNN to handle the A-SGNS word embedding vectors resulted in its enhanced performance in a dynamic web environment [51], with an average accuracy of 0.813.

TABLE II. Precision, Recall, F1-score and Accuracy with 350,000 Training Samples

| Algorithm | Precision | | Recall | | F1-score | | Accuracy |
|---|---|---|---|---|---|---|---|
| | Class 1 | Class 0 | Class 1 | Class 0 | Class 1 | Class 0 | |
| SVM + TF-IDF | 0.50 | 0.51 | 0.62 | 0.39 | 0.55 | 0.44 | 0.623 |
| NB + TF-IDF | 0.50 | 0.513 | 0.626 | 0.39 | 0.55 | 0.443 | 0.62 |
| ANN + TF-IDF | 0.5 | 0.50 | 0.42 | 0.58 | 0.45 | 0.53 | 0.70 |
| RNN + A-SGNS | 0.62 | 0.57 | 0.45 | 0.73 | 0.52 | 0.64 | 0.813 |

### C. Performance Evaluation of Crawling Phase

#### 1. Performance Metrics

The performance of the six focused crawlers were measured by using harvest rate and irrelevance ratio can be shown in the equations (38) and (39).

#### 2. Harvest Rate

Harvest rate is defined as the ratio of the number of relevant web pages downloaded out of total number of web pages downloaded. The harvest rate *(hr)* can be formulated as follows in equation (38).

$$hr = \frac{R_{wp}}{N_{wp}} \tag{38}$$

where hr is the harvest Rate, $R_{wp}$ is the number relevant web pages downloaded, and $N_{wp}$ is the total number of web pages downloaded.

### 3. Irrelevance Ratio

Irrelevance ratio is defined as the ratio of number of irrelevant web pages downloaded out of total number of web pages downloaded. The irrelevance ratio can be formulated as follows in equation (39).

$$ir = \frac{r_i \cap n_i}{n_i} \quad (39)$$

where ir is the irrelevance ratio, $r_j$ is the number of relevant web pages downloaded, and $n_j$ is the total number of web pages downloaded.

## D. Analysis of Crawling Phase

The experimental results were evaluated for all the four focused crawlers, namely, the SVM + TF-IDF, NB + TF-IDF, ANN + TF-IDF and the proposed RNN + A-SGNS. For the NB, SVM, and ANN, the TF-IDF-based cosine similarity was given as an input feature, while for the RNN, the SGNS-based cosine similarity was the input feature.

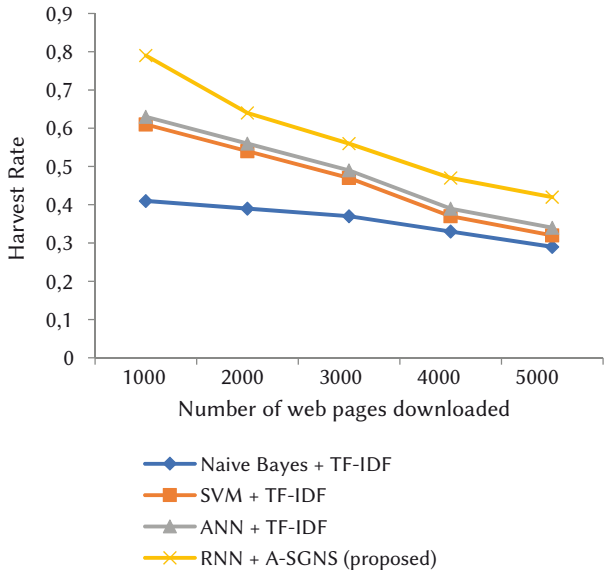



Fig. 4. Average harvest rate for ten topics for the SVM + TF-IDF, NB + TF-IDF, ANN + TF-IDF and RNN + A-SGNS crawlers.

Fig. 4 shows the average harvest rate and Fig. 5 shows the average irrelevance ratio of the SVM + TF-IDF, NB + TF-IDF, ANN + TF-IDF and RNN + SGNS crawlers, respectively. The TF-IDF-based features consider similarity only if the topic term co-occurs on the web page. As a result, the SVM + TF-IDF, NB + TF-IDF crawler, and ANN + TF-IDF crawler considers most web pages that are semantically related to the topic as irrelevant. The SVM+TF-IDF, NB + TF-IDF and ANN + TF-IDF crawlers produced an average harvest rate of 0.32, 029 and 0.34, along with a high irrelevance ratio of 0.68, 0.71 and 0.66,respectively. The A-SGNS is a context learning-based algorithm that considers the semantic relatedness between the topic and the web page term. Owing to this advantage, it considers the semantically related web page as a relevant web page, and produced an average harvest rate of 0.42 and a low irrelevance ratio of 0.58, thus outperforming the other focused SVM + TF-IDF, NB+ TF-IDF and ANN + TF-IDF crawlers.

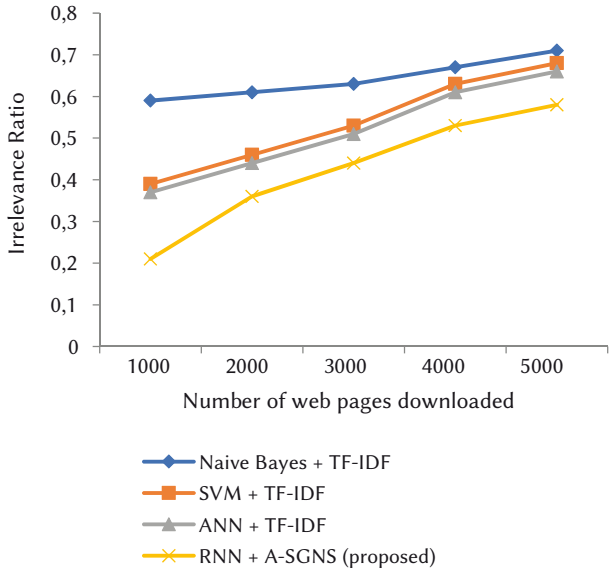



Fig. 5. Average irrelevance ratio for ten topics for the SVM + TF-IDF, NB + TF-IDF, ANN + TF-IDF and RNN + A-SGNS crawlers.

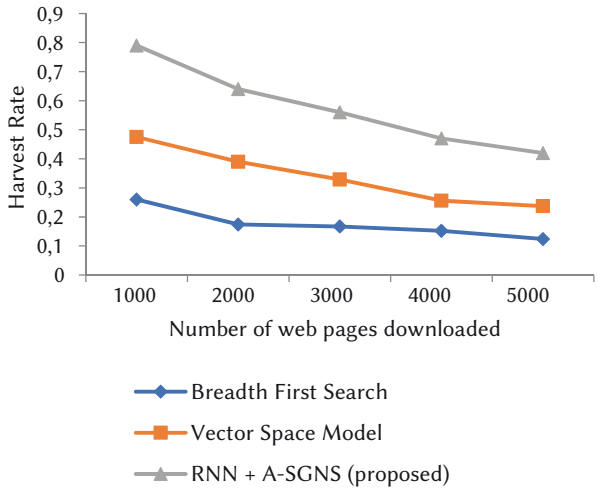



Fig. 6. Average harvest rate of ten topics for the BFS, VSM and RNN+A-SGNS crawlers.

To retrieve the associated web pages without determining their topical preferences, the breadth-first crawler explicitly selects unvisited hyperlinks. The VSM makes use of the TF-IDF to compute topical similarities but failed to capture the semantic similarity. As a result, the average harvest rate of the BFS and VSM is less than that of the RNN + SGNS and the average irrelevance ratio of the BFS and VSM is higher than that of the RNN + SGNS. Fig. 6 and Fig. 7 show the average harvest rate and average irrelevance ratio of the BFS, VSM and RNN+A-SGNS respectively. Right from the beginning, the BFS starts retrieving irrelevant results, and after 5000 web page crawls produced an average harvest rate of 0.124 and an irrelevance ratio of 0.876. The VSM crawler performed better than the BFS crawler because of the relevance computation. The VSM crawler makes use of the TF-IDF to compute topical similarities but failed to capture the semantic similarity. After 5000 web page crawls, the VSM crawler produced an average harvest rate of 0.237 and an irrelevance ratio of 0.763. The proposed RNN+A-SGNS crawler outperformed both the BFS and VSM crawlers with an average harvest rate of 0.42 and an irrelevance ratio of 0.58 after 5000 web page crawls.

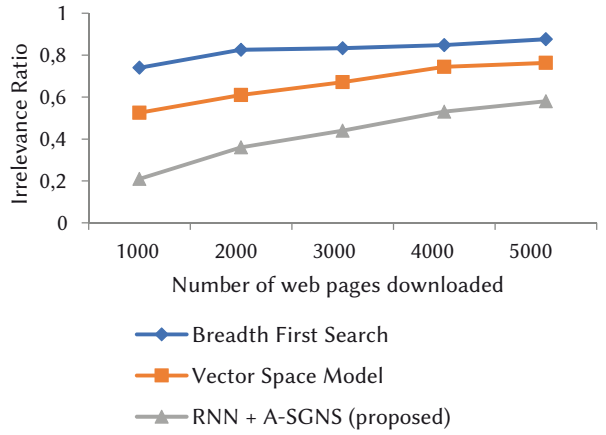



Fig. 7. Average irrelevance ratio of ten topics for the BFS, VSM and RNN+A-SGNS crawlers.

Ontology learning-based crawlers use a domain-specific ontology to ascertain the topical similarity between a topic and web pages. An ontology is a well-known representation that helps find semantic similarity. Ontologies are domain-specific and designed by domain experts. A human error in ontology design results in the retrieval of wrong results. In this work, WordNet ontology [52] for the semantic representation of words was used in the design of the optimized Naive Bayes (ONB) crawler, the ontology learning-based crawler using the ANN (OL-ANN), and the semi- supervised learning-based crawler using the SVM (SSL-SVM). Ontology learning on the dynamic internet is a difficult and time-consuming process. Given the limitations of ontologies and ontology learning, these crawlers performed poorly on the dynamic internet in terms of the harvest rate and irrelevance ratio, when compared to the proposed methodology. The ONB, ontology learning-based crawler using the ANN, the semi-supervised learning-based crawler using the SVM, and the proposed crawler produced an average harvest rate of 0.39, 0.37,0.36 and 0.42, respectively, and an average irrelevance ratio of 0.61, 0.63, 0.64 and 0.58, respectively. This clearly shows that the proposed crawler outperformed the ontology learning-based crawler. Fig. 8 and Fig. 9 show a comparison of the results of the ONB, OL-ANN, SSL-SVM and the proposed crawler in terms of the harvest rate and irrelevance ratio, respectively.

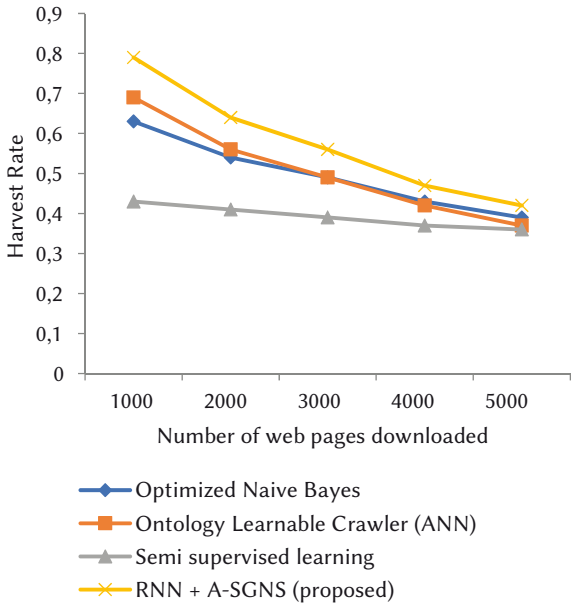



Fig. 8. Average Harvest Rate of ten topics for ONB, OL-ANN, SSL-SVM and RNN+A-SGNS.

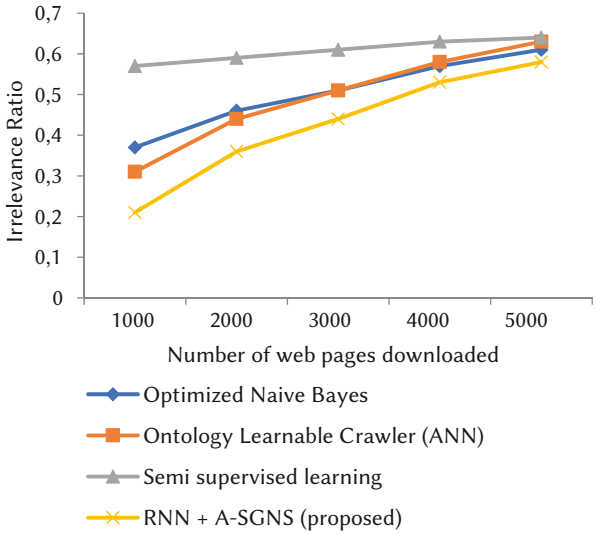



Fig. 9. Average Irrelevance Ratio of ten topics for ONB, OL-ANN, SSL-SVM and RNN+A-SGNS.

## VI. Conclusion and Future Work

The A-SGNS model presented here was intended to optimize the performance of the focused web crawler. This work considers both the syntactic and semantic similarity between the topic and web page documents. The model first computes the A-SGNS model, from which the cosine similarity of the topic and document terms is calculated. The similarity vectors are given as input to the recurrent neural network to classify the web page, based on its relevance. The results of the experiment have demonstrated that the proposed system has increased the efficiency of the focused crawler, outperforming the breadth-first, VSM, and TF-IDF-based learning crawlers as well as those based on ontology learning. In conclusion, the proposed method is ideally suited to focused crawlers and has conclusively proved its efficacy.

Future directions include plans for the design of a crawler using long short-term memory networks (LSTM) or the gated recurrent unit (GRU) to resolve the long-term dependency problem of the RNN in learning sequences, brought on by problems with the vanishing gradient.

## References

[1] "Internet Live Status," 2020. [Online]. Available: https://www.internetlivestats.com/total-number-of-websites/.

[2] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine BT - Computer Networks and ISDN Systems," *Comput. Networks ISDN Syst.*, vol. 30, no. 1–7, pp. 107–117, 1998.

[3] G. Ayers, J. H. Ahn, C. Kozyrakis, and P. Ranganathan, "Memory Hierarchy for Web Search," *Proc. - Int. Symp. High-Performance Comput. Archit.*, vol. 2018-Febru, pp. 643–656, 2018.

[4] Auf Wiedersehen, "The Architecture of a Large-Scale Web Search Engine, circa 2019," 2019. [Online]. Available: https://0x65.dev/blog/2019-12-14/the-architecture-of-a-large-scale-web-search-engine-circa-2019.html.

[5] B. Muller, "How search engines work: Crawling, Indexing, and Ranking," *Moz Pro*, 2020. [Online]. Available: https://moz.com/beginners-guide-to-seo/how-search-engines-operate.

[6] A. Hliaoutakis, G. Varelas, E. Voutsakis, E. G. M. Petrakis, and E. Milios, "Information Retrieval by Semantic Similarity," *Int. J. Semant. Web Inf. Syst.*, vol. 2, no. 3, pp. 55–73, 2011.

[7] Z. Liu, Y. Du, and Y. Zhao, "Focused Crawler Based on Domain Ontology and FCA," *J. Inf. Comput. Sci.*, vol. 8, no. 10, pp. 1909–1917, 2011.

[8] Z. Geng, D. Shang, Q. Zhu, Q. Wu, and Y. Han, "Research on improved focused crawler and its application in food safety public opinion analysis," *2017 Chinese Autom. Congr.*, pp. 2847–2852, 2017.

[9] T. Hassan, C. Cruz, and A. Bertaux, "Predictive and evolutive cross-referencing for web textual sources," *Proc. Comput. Conf. 2017*, vol. 2018-Janua, no. July, pp. 1114–1122, 2018.

[10] S. Chakrabarti, M. van den Berg, and B. Dom, "Focused crawling: a new approach to top-specific Web source discovery," *Comput. Networks*, vol. 31, no. 11–16, pp. 1623–1640, 1999.

[11] F. Menczer, F. Menczer, G. Pant, G. Pant, P. Srinivasan, and P. Srinivasan, "Topical Web Crawlers: Evaluating Adaptive Algorithms," *ACM Trans. Internet Technol.*, vol. V, no. February, p. 38, 2003.

[12] J. R. Park, C. Yang, Y. Tosaka, Q. Ping, and H. El Mimouni, "Developing an automatic crawling system for populating a digital repository of professional development resources: A pilot study," *J. Electron. Resour. Librariansh.*, vol. 28, no. 2, pp. 63–72, 2016.

[13] G. H. Agre and N. V. Mahajan, "Keyword focused web crawler," *2nd Int. Conf. Electron. Commun. Syst. ICECS 2015*, pp. 1089–1092, 2015.

[14] Y. Du, W. Liu, X. Lv, and G. Peng, "An improved focused crawler based on Semantic Similarity Vector Space Model," *Appl. Soft Comput. J.*, vol. 36, pp. 392–407, 2015.

[15] G. Salton, A. Wong, and C. Yang, "Information Retrieval and Language Processing: A Vector Space Model for Automatic Indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613-620, 1975.

[16] P. Bedi, A. Thukral, and H. Banati, "Focused crawling of tagged web resources using ontology," *Comput. Electr. Eng.*, vol. 39, no. 2, pp. 613–628, 2013.

[17] A. I. Saleh, A. E. Abulwafa, and M. F. Al Rahmawy, "A web page distillation strategy for efficient focused crawling based on optimized Naïve bayes (ONB) classifier," *Appl. Soft Comput. J.*, vol. 53, pp. 181–204, 2017.

[18] H. D. and F. K. Hussain, "SOF: a semi-supervised ontology-learning-based focused crawler," *Concurr. Comput. Pract. Exp.*, vol. 25, no. 6, pp. 1755–1770, 2013.

[19] H. T. Zheng, B. Y. Kang, and H. G. Kim, "An ontology-based approach to learnable focused crawling," *Inf. Sci. (Ny).*, vol. 178, no. 23, pp. 4512–4522, 2008.

[20] A. Capuano, A. M. Rinaldi, and C. Russo, "An ontology-driven multimedia focused crawler based on linked open data and deep learning techniques," *Multimed. Tools Appl.*, 2019.

[21] Y. Li, Z. A. Bandar, and D. McLean, "An approach for measuring semantic similarity between words using multiple information sources," *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 4, pp. 871–882, 2003.

[22] J. Hosseinkhani, H. Taherdoost, and S. Keikhaee, "ANTON Framework Based on Semantic Focused Crawler to Support Web Crime Mining Using SVM," *Ann. Data Sci.*, 2019.

[23] D. Mukhopadhyay and S. Sinha, "Domain-Specific Crawler Design," pp. 85–112, 2019.

[24] J. Qiu, Q. Du, W. Wang, K. Yin, C. Lin, and C. Qian, "Topic Crawler for OpenStack QA Knowledge Base," *Proc. - 2017 Int. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discov. CyberC 2017*, vol. 2018-Janua, pp. 309–317, 2018.

[25] T. Suebchua, B. Manaskasemsak, A. Rungsawang, and H. Yamana, "History-enhanced focused website segment crawler," *Int. Conf. Inf. Netw.*, vol. 2018-Janua, pp. 80–85, 2018.

[26] G. Xu, P. Jiang, C. Ma, and M. Daneshmand, "A Focused Crawler Model Based on Mutation Improving Particle Swarm Optimization Algorithm," *Proc. - 2018 IEEE Int. Conf. Ind. Internet, ICII 2018*, no. Icii, pp. 173–174, 2018.

[27] H. Dong and F. K. Hussain, "Self-adaptive semantic focused crawler for mining services information discovery," *IEEE Trans. Ind. Informatics*, vol. 10, no. 2, pp. 1616–1626, 2014.

[28] W. J. Liu and Y. J. Du, "A novel focused crawler based on cell-like membrane computing optimization algorithm," *Neurocomputing*, vol. 123, pp. 266–280, 2014.

[29] P. Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy," vol. 1, 1995.

[30] "Natural Language Processing Tool Kit (NLTK)," 2020. [Online]. Available: https://www.nltk.org/.

[31] E. L. and E. K. Bird, Steven, *Natural Language Processing with Python*. O'Reilly Media Inc, 2009.

[32] H. Palangi *et al.*, "Deep Sentence embedding using long short-term memory networks: Analysis and application to information retrieval," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 4, pp. 694–707, 2016.

[33] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 4, no. January, pp. 3104–3112, 2014.

[34] T. Mikolov, M. Karafiát, L. Burget, C. Jan, and S. Khudanpur, "Recurrent neural network based language model," *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH 2010*, no. September, pp. 1045–1048, 2010.

[35] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations ofwords and phrases and their compositionality," *Adv. Neural Inf. Process. Syst.*, pp. 1–9, 2013.

[36] Y. Goldberg and O. Levy, "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method," no. 2, pp. 1–5, 2014.

[37] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *1st Int. Conf. Learn. Represent. ICLR 2013 - Work. Track Proc.*, pp. 1–12, 2013.

[38] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *COLT 2010 - 23rd Conf. Learn. Theory*, pp. 257–269, 2010.

[39] S. Ruder, "An overview of gradient descent optimization algorithms," pp. 1–14, 2016.

[40] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Inf.*, vol. 10, no. 4, pp. 1–68, 2019.

[41] G. Chen, "A Gentle Tutorial of Recurrent Neural Network with Error Backpropagation," pp. 1–10.

[42] G. Hinton, N. Srivastava, and K. Swersky, "Lecture 6a Overview of mini-batch gradient descent," *Neural Networks Mach. Learn. Coursera*, 2012.

[43] G. van Rossum, "Python tutorial, Technical Report CS-R9526," *Cent. voor Wiskd. en Inform. (CWI), Amsterdam*, 1995.

[44] "Python 3.6," 2016. [Online]. Available: https://www.python.org/downloads/release/python-360/.

[45] "Spyder IDE," 2009. [Online]. Available: https://www.spyder-ide.org/.

[46] L. Richardson, "Beautiful Soup Documentation Release 4.4.0." 2019.

[47] "urllib," *Python*, 2020. [Online]. Available: https://docs.python.org/3/library/urllib.html.

[48] "lxml parser," *Python*, 2020. [Online]. Available: https://lxml.de/elementsoup.html.

[49] G. Varoquaux, L. Buitinck, G. Louppe, O. Grisel, F. Pedregosa, and A. Mueller, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 1, pp. 2825–2830, 2011.

[50] D. Isa, L. H. Lee, V. P. Kallimani, and R. Rajkumar, "Text document preprocessing with the bayes formula for classification using the support vector machine," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 9, pp. 1264–1272, 2008.

[51] H. Salehinejad, S. Sankar, J. Barfett, E. Colak, and S. Valaee, "Recent Advances in Recurrent Neural Networks," pp. 1–21, 2017.

[52] Princeton University, *"About WordNet." WordNet.* Princeton University, 2010.



P. R. Joe Dhanith

P.R.Joe Dhanith received his B.Tech degree in Information Technology from Anna University in 2010 and M.E degree in Computer Science and Engineering from Anna University in 2012. He is currently pursuing his Ph.D degree in Computer Science and Engineering at National Institute of Technology Puducherry. His main research interests includes web mining, web crawling and information retrieval.



B.Surendiran

B. Surendiran is currently working as an Assistant Professor in the Department of Computer Science and Engineering at National Institute of Technology Puducherry, Karaikal, India. He completed his Ph.D in Computer Science and Engineering at National Institute of Technology Tiruchirapalli. His research interest includes recommender systems, machine learning and data mining. He has published more than 30 papers in international journals.

S.P.Raja

S. P. Raja completed his B. Tech in Information Technology in the year 2007 from Dr. Sivanthi Aditanar College of Engineering, Tiruchendur. He completed his M.E. in Computer Science and Engineering in the year 2010 from Manonmaniam Sundaranar University, Tirunelveli. He completed his Ph.D. in the year 2016 in the area of Image processing from Manonmaniam Sundaranar University, Tirunelveli. His area of interest is image processing and cryptography. He is having more than 13 years of teaching experience in engineering colleges. Currently he is working as an Associate Professor in the department of Computer Science and Engineering in Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai. He published 30 papers in International Journals, 24 in International conferences and 12 in national conferences. He is an Editorial Board Member of International Journal of Interactive Multimedia and Artificial Intelligence.

# BILROST: Handling Actuators of the Internet of Things through Tweets on Twitter using a Domain-Specific Language

Daniel Meana-Llorián, Cristian González García*, B. Cristina Pelayo G-Bustelo, Juan Manuel Cueva Lovelle

MDE Research Group, Department of Computer Science, University of Oviedo, Oviedo, Asturias (Spain)

Received 31 December 2019 | Accepted 16 January 2021 | Published 31 January 2021

## Abstract

In recent years, many investigations have appeared that combine the Internet of Things and Social Networks. Some of them addressed the interconnection of objects as Social Networks interconnect people, and others addressed the connection between objects and people. However, they usually used interfaces created for that purpose instead of using familiar interfaces for users. Why not integrate Smart Objects in traditional Social Networks? Why not control Smart Objects through natural interactions in Social Networks? The goal of this paper is to make easier to create applications that allow non-experts users to control Smart Objects actuators through Social Networks through the proposal of a novel approach to connect objects and people using Social Networks. This proposal will address how to use Twitter so that objects could perform actions based on Twitter users' posts. Moreover, it will be presented a Domain-Specific language that could help in the task of defining the actions that objects could perform when people publish specific content on Twitter.

## Keywords

Internet Of Things, Smart Objects, Model-Driven Engineering, Domain-Specific Language, Social Networks, Twitter.

DOI: 10.9781/ijimai.2021.01.004

## I. Introduction

THE Internet of Things (IoT) is a term that has gained popularity in recent years among common people due to the wishes for interconnecting the whole things around them. They want to connect objects located at home to the Internet like the fridge, the oven, and so on, to manage them or know real-time information about them. Some examples of their expectations can be a fridge with the capability of alerting them when a product is running out, or an oven capable of turning on when they are arriving home. Moreover, the rise of mobile devices like smartphones, tablets, wearables, or any other devices connected to the internet like sensors, smart tags, and so on, has contributed to the popularity of the IoT.

Despite the growing popularity of the IoT, it is not so present in people's lives as was expected. What are the causes? The answer could be the complexity of managing the Smart Objects. The Smart Objects are usually composed of other objects without intelligence like sensors and actuators [1], also called Not-Smart Objects. On the one hand, managing Smart Objects' actuators could be complex: what actions to do, when the actuator must work, or how it must do it. On the other side, recollecting and interpreting data from sensors could also be so complex. Moreover, establishing connections with these Smart Objects is also an advanced task. Because of that, the goal of this work is to reduce the complexity of developing a specific type of application that Smart Objects could run. It aims to facilitate the creation of applications for users without advanced programming knowledge (hereinafter non-expert users) that allow them to control Smart Objects' actuators through a novel way of communication, the traditional Social Networks like Twitter.

The use of Social Networks in communication with the Smart Objects has several advantages over other commons solutions like architectures client-server over HTTP protocol. For instance, this approach does not require using a specific application to intercommunicate users with objects. Users could use any application that allows them to use the chosen Social Network. Besides, this solution may be more intuitive for people who use Social Networks in their day-to-day lives.

Therefore, the hypothesis is the next: **It is possible to facilitate the creation of applications that allow non-experts users to control Smart Objects' actuators through Social Networks for humans.**

To achieve this goal, it is proposed the creation of a Domain-Specific Language (DSL) applying Model-Driven Engineering (MDE), that have been called Bilrost Specific Language (BSL). The BSL was designed focused on the ease of use and it provides the required features to enable users to define the rules and properties needed to set up the Smart Objects' actuators with their actions and to communicate Smart Objects and users through Social Networks. This proposal is capable of generating application projects for Smart Objects where the whole logic needed to communicate the Smart Objects with users through Social Networks is already implemented. Thus, non-expert users only need to implement the logic needed to manage the Smart Objects' actuators according to a skeleton available in the generated projects.

* Corresponding author.
E-mail address: gonzalezcristian@uniovi.es

In other words, non-expert users do not need to know how to connect Smart Objects to Social Networks, they only need a basic knowledge of programming applications for specific Smart Objects.

This paper address the first part of the research idea presented in [2]. Bilrost not only will address the controlling of actuators through Twitter users' posts, but also will address the communication Object-Object and Object-Human. In the final stages, Bilrost will enable Smart Objects to post messages on Twitter that will invoke the actions of other Smart Objects, and they will be able to share their status with users. Finally, Bilrost will also be able to generate Smart Objects without the necessity of programming skills, achieving the automate completion of the skeletons that are going to be presented in this paper.

In the following lines, the proposal is going to be presented (Section II), addressing what it is, how it works, how its architecture is, how the BSL is. Next, the proposal is going to be evaluated by comparing opinions from two different user profiles after completing an assigned task (Section III). After present and evaluate the proposal, the next section is going to address the related work (Section IV), present the conclusions (Section V), and describe the possible future work that can be done from here (Section VI).

## II. Case Study

This section is going to address Bilrost. It was developed to investigate if the communication between objects and people through traditional Social Networks like Twitter is possible.

Bilrost aims to enable non-expert users to generate applications that easily connect people and objects. The goal is to achieve that anybody, without knowledge about how to connect devices to the Internet, will be able to generate applications that connect their devices with them through Social Networks like Twitter. However, to use this proposal, it is required a basic programming knowledge, hence, the goal of Bilrost is to help people with that basic programming knowledge, or non-expert users, to connect their Smart Objects to Social Networks to perform actions according to the messages that their owners sent to the Social Network. For instance, this proposal is suitable for users who have specific knowledge about developing simple applications for a Raspberry Pi, but they do not have enough programming knowledge to develop complex applications that use the Twitter's API.

Twitter was the Social Network chosen due to some interesting features like the lack of reciprocation in the relationships and its mark-up language of tweets (hashtags, mentions, etc.). However, Bilrost is prepared to use more Social Networks in the future.

In short, the main aim of Bilrost is that everybody can handle Smart Objects remotely without specific knowledge. For that, Bilrost generates a skeleton of an application where all the logic needed to connect Smart Objects to Twitter is already implemented, but users must complete it with the logic required for using the actuators of each device.

### A. Work-Cycle

Bilrost enables non-expert users to develop applications that handle their Smart Objects' actuators through Twitter. These non-expert users must interact with Bilrost twice to obtain a final application that enables them to handle their Smart Objects through tweets. Thus, the work cycle consists of two steps: **project generation** and **project completion**.

It is important to mention that Bilrost does not generate all logic, but it generates the logic required to connect Smart Objects to Twitter and it creates a skeleton already prepared to be completed with the specific logic of each Smart Object. This skeleton contains empty methods that users must fill as they want with the logic to perform actions. These methods will be called according to the received tweets.

#### 1. Projects Generation

The work cycle starts with the generation of application projects. Firstly, users must write the definition of a device using the BSL syntax (the syntax will be explained in the next section). After writing the definition of the device, the project generator processes the definition to generate an application project where the logic required to establish communication with Twitter is already implemented. Moreover, this application project contains a skeleton that makes easier the specific implementation for each action.

In short, this step consists of two sub-steps:

1. Writing the definition of a Smart Object using the BSL syntax.
2. Using Bilrost to generate the skeleton of the application from the definition written in the previous step.

#### 2. Projects Completion

The work cycle ends with the completion of projects generated in the previous step. The generated projects contain empty methods which users must fill with the logic needed to enable actuators to perform actions.

This step requires basic knowledge about developing applications for Smart Object because users must implement the actions that an actuator can perform. Bilrost does not take part in this step because the proposal is focused on the generation of the logic needed to connect people and Smart Objects through Social Networks. The automation of this step is a part of the future work to take into consideration in future research.

After filling the skeleton, the project is ready to deploy in the target Smart Object.

In short, this step consists of the other two steps:

1. Implement the specific logic to control the Smart Object. Users must implement Smart Object's actions which will be triggered by tweets.
2. Deploy the final application in the Smart Object.

### B. Architecture

The architecture of Bilrost can be divided into two components: the **BSL Parser**, and the **Project Generator**.

The first component, the BSL Parser, is responsible for processing the definition of Smart Objects written using the BSL syntax. The result obtained in this component is sent to the second component, the Project Generator, which takes the result processes it to generate a project ready to connect a Smart Object to Twitter. The generated project contains the skeleton that users must fill as it was already said. When users complete the implementation, the program will be finished, and it will be ready to be deployed into devices like Android, Raspberry Pi, or other devices supported by BSL.

#### 1. Bilrost-Specific Language Parser

The BSL Parser is the component responsible for processing definitions written using BSL. It receives a file, written with the BSL syntax, with the definition of a device and generates a tree which contains all required data. This tree will be sent to Project Generator in JSON format.

The content of the file is a model that represents a device with its actuators, the actuators' actions, the Social Networks that it will use with the needed parameters, and more required information that will be explained later.

For instance, Code 1 shows a little example of a device definition using BSL. This definition would be the input of the BSL parser and represents a device whose programming language is Python and has two actuators: a LED and a screen. The LED's actions are '*on*'

and 'off', and the screen's action is 'show'. This device will establish communication with Twitter but only two users will be able to handle it: 'dani_meana' and 'bilrost_bridge'. The rest of the fields are going to be explained in Section II.C.

```
DEVICE IN PYTHON
  FILTER BY 'bilrost', 'uniovi'
  SOCIAL NETWORKS
    CONNECT TO TWITTER
    USERNAME 'username'
    PASSWORD 'password'
    USERS 'dani_meana', 'bilrost_brdige'
  ACTUATORS
    DEFINE 'led'
      LOCATION 'rips'
      ACTIONS 'on', 'off'
    DEFINE 'screen'
      LOCATION 'rpi'
      ACTIONS 'show'
```

Code 1. Definition of a device using Bilrost-Specific Language.

### 2. Project Generator

The Project Generator is the component responsible for generating projects that already contain the logic needed to establish communications with Social Networks and the skeleton that users must fill.

The Project Generator waits for the tree in JSON format generated in the previous component to generate the application. From the tree, the generator chooses a template that fits the input data and fills it with the data of the tree. The communication with Social Networks and the processing of the data is already implemented so users only must address the concrete implementation of each action.

## C. Bilrost-Specific Language

A textual DSL for Bilrost called Bilrost-Specific Language was also designed. Code 2 shows the BSL's context-free grammar written in *Backus-Naur Form* (BNF) although there are tokens that are not explained as WORD or COMMA because they are part of the lexical step. Most tokens were defined to represents the same word as their names except WORD and COMMA. The token WORD represents any string composed of letters, numbers, underscore (_), and dash (-). The token COMMA represents the character comma (,).

BSL does not distinguish between lower case and uppercase, allows writing all code in a single line or multiple lines mixing uppercase and lower case, and changing the order of the different blocks that define a device. Each file written in BSL defines a unique device, hence, users must create as many files as devices they want to define. The definition of a device is composed of the properties of the device and two other different blocks: Social Networks, and Actuators.

To write a program with BSL users must start defining the project language that they want to generate and after that, they must write the properties of the device like filters, Social Networks to connect with their properties, and the actuators.

Furthermore, there are also comments in BSL. The syntax of comments is the same as Python syntax. It starts with a hash sign (#) and ends with a new line.

```
<device>            ::= DEVICE IN <platform> <properties> END
<platform>          ::= PYTHON
                      | JAVA
                      | ANDROID
<properties>        ::= <property>
                      | <properties> <property>
<property>          ::= <filter>
                      | <social-networks>
                      | <actuators>
<filter>            ::= FILTER BY <filters>
<filters>                   ::= WORD
                      | WORD COMMA <filters>
<social-networks>   ::= SOCIAL NETWORKS <social-networks-list>
<social-networks-list>      ::= <social-network>
                              | <social-networks-list> <social-network>
<social-network>    ::= CONNECT TO TWITTER <twitter-properties>
<twitter-properties>        ::= <username> <password> <users>
                      | <username> <users> <password>
                      | <password> <username> <users>
                      | <password> <users> <username>
                      | <users> <username> <password>
                      | <users> <password> <username>
                      | <username> <password>
                      | <password> <username>
<username>          ::= USERNAME WORD
<password>          ::= PASSWORD WORD
<users>             ::= ALLOW <users-list>
<users-list>        ::= WORD
                      | WORD COMMA <users-list>
<actuators>         ::= ACTUATORS <actuators-list>
<actuators-list>    ::= <actuator>
                      | <actuators-list> <actuator>
<actuator>          ::= DEFINE WORD <location> ACTIONS <actions>
                    | DEFINE WORD ACTIONS <actions> <location>
<location>          ::= LOCATION WORD
<actions>           ::= WORD
                        | <actions> COMMA WORD
```

Code 2. Context-free grammar in BNF.

### 1. Device Definition

The first step is to define the device's properties. The first property that users must define is the application language. This proposal can generate application projects in Python, Java, and Android.

Another property of a device is the filters. Users must define some keywords that help to identify the device in Social Networks. These keywords can be used to filter the messages that the device search in Social Networks. A device will only perform actions if the messages which will arrive contain the specified filters. A device can have as many filters as users want but at least one.

The next code is the skeleton to define a device.

```
DEVICE IN PYTHON | JAVA | ANDROID
  FILTER BY ...
  SOCIAL NETWORKS ...
  ACTUATORS ...
```

The next code is the skeleton to indicate filters.

```
FILTER BY 'filter1', 'filter2', ...
```

After that, users must indicate the Social Networks that they want to connect their device and the actuators that composed the device.

### 2. Social Networks

This prototype was developed to work with Twitter in the first stage, but it is adaptable to other Social Networks. Hence, BSL allows the definition of several Social Networks. The block to indicate the Social Networks that the device will use starts with the reserved words SOCIAL NETWORKS followed by the parameters required by each Social Network.

The next code is the skeleton to indicate which Social Networks will use the device.

```
SOCIAL NETWORKS
  CONNECT TO TWITTER | OTHERS
```

As mentioned above, this prototype uses Twitter as the Social Network, so the parameters that users must write are their credentials. For that, there are two reserved words USERNAME and PASSWORD. The Project Generator uses these parameters to obtain the tokens required by Twitter API, hence, the final application will not contain the credentials.

Furthermore, there is an optional third parameter to control what users can call the actions of an actuator. It adds a security filter avoiding the control of users' actuators by malicious users. For that, users must use the reserved word ALLOW.

The next code shows the skeleton to configure Twitter as Social Network.

```
CONNECT TO TWITTER
  USERNAME 'username'
  PASSWORD 'password'
  ALLOW 'user1', 'user2', ...
```

### 3. Actuators

Bilrost uses Twitter to invoke the actions of users' actuators. For that, users use BSL to define the devices' actuators with their actions. The block needed to define the actuators starts with the reserved word ACTUATORS and to define each actuator users must write the reserved word DEFINE followed by the name that they want to assign to the actuator. Furthermore, an actuator has several properties that users must define. These properties are the location of the actuator and the name of the actuator's actions.

The location is useful to filter the messages that arrive at the device. In this way, the device could receive messages with a specific location and only the actuators in this location would respond. To specify a location, users must write the reserved word LOCATION.

The name of the actuator's actions is used to enable the invocation of the actuator's actions through Social Networks. Moreover, the name of actions is also the name of the methods that users must fill in the project competition step. A device can have as many actions as users want but at least one.

The next code shows the skeleton to indicate the actuators that compose the device.

```
ACTUATORS
  DEFINE 'name'
    LOCATION 'filter1'
    ACTIONS 'action1', 'action2', ...
```

### D. Communication Through Twitter

The communication through Twitter is made by tweeting in the timeline. A tweet to control an actuator must contain the device's filters, the actuator's filters, the actuator's name, the action to call, and the parameters that the action could need. Moreover, due to the Twitter limitations (repeated tweets), the messages should contain more content at the final of the message, for instance, the timestamp

To represent the filters of a device, its location, and/or its name, users must use hashtags (#) whereas the action to call must be plain text and its parameters must be enclosed in quotes. Furthermore, users must implement how to parse the parameters in the project completion step.

The hashtag that represents the name is the unique one that is not mandatory. If the name was not specified, all actuators which are in the location would execute the action specified.

In the following lines, there are examples of tweets that handle actuators:

- **#bilrost #uniovi #rpi #red on**: It invokes the action named *on* of an actuator named *red*, located in *rpi*, and it is filtered by the keywords *bilrost* and *uniovi*.
- **#bilrost #uniovi #smartphone #flash on**: It invokes the action named *on* of an actuator named *flash*, located in *smartphone*, and it is filtered by the keywords *bilrost* and *uniovi*.
- **#bilrost #uniovi #rpi off**: It invokes the action named *off* of all actuators located in *rpi*, and they are filtered by the keyword *bilrost* and *uniovi*.

The next examples show the use of actions parameters.

- **#bilrost #uniovi #rpi #display show "hello world"**: It invokes the action named *show* of an actuator named *display*, located in *rpi*, and it is filtered by the keywords *bilrost* and *uniovi*. Moreover, it sends the parameter *hello world* to the action.
- **#bilrost #uniovi #lab #thermostat set "22"**: It invokes the action named *set* of an actuator named *thermostat*, located in *lab*, and they are filtered by the keywords *bilrost* and *uniovi*. Moreover, it sends the parameter *22* to the action.

The filters, the location, and the name of an actuator are the same type of words so the generated application will check all possible combinations that can fit with the defined device.

## III. Evaluation and Discussion

This section is going to describe the evaluation process and discuss the obtained results.

### A. Methodology

The evaluation process consists of two phases where information was collected to check if the proposal is useful not only for expert users but also for users that have not knowledge about the IoT. In the first phase, it is measured the time that users spent to complete a specific task. After that, in the second phase, users filled a survey based on the Likert scale to measure their opinions about the proposal and its usefulness.

### 1. Phase 1

In this phase, users of two different profiles had to complete a task that emulated a real scenario that required two actuators connected to Twitter with several actions. The entire task is in the following paragraph.

The required task consisted of defining a device which had two actuators: a fan and a thermostat. The device was a Raspberry Pi and the programming language for the development was Python. The

device would have to look for messages that contained the keywords bilrost, evaluation, and test, the device would have to be connected to Twitter and the only user who would be capable of handling the device would have to be *@bilrost_bridge*. The fan's actions were to turn it on, turn it off, and set its speed. The thermostat's actions were to turn it on, turn it off, increase the temperature, and decrease the temperature. The location of both actuators would have to be the same. When users would have finished the task, they would have to identify the tweets that perform the next actions:

- Turning the fan on.
- Turning the thermostatic on.
- Increasing the temperature.
- Setting the fan speed to 2000 rpm.

Different users were chosen between two profiles, people who had knowledge about the IoT and people without this knowledge because the target of using a DSL is to avoid the requirement of knowledge about a specific technology. A total of 20 participants took part in the evaluation process: 13 participants with knowledge about the IoT and 7 participants without this knowledge but all of them with basic programming knowledge.

During the evaluation, every user had the documentation needed to perform the task where the objectives and the BSL syntax were explained, and some examples were also available to make it easier to understand how the system works. Furthermore, they had time to read the documentation without a limit of time with the possibility of asking any doubt about the system. After that, the system was shown to them and they had more time to test it in order to try to remove learning effects from the gathered results. When users were ready, the task was explained and gave them more time to understand it and think about how to solve them but without access to the system.

Finally, when the participants had said that they are ready and they understood everything, the measurement of times started, and it stopped when users completed the task correctly.

The time limit to complete the task was four times greater than the time spent by the developer of the prototype. It is important to mention that the participants had the documentation available to consult during the evaluation process.

### 2. Phase 2

After finishing the first phase, users must complete an anonymous survey about this proposal. To create the survey, the 5-points Likert Scale was used because it is the most used in the design of scales. The given options were the following: 1 as strongly disagree, 2 as disagree, 3 as neutral, 4 as agree, and 5 as strongly agree.

The survey was composed of ten declarations that ask users for their opinions about the creation of applications that interconnect objects and humans using this proposal and its possible impact on the IoT.

The survey is composed of a set of ten declarations that are shown in Table I.

TABLE I. SURVEY GIVEN TO THE USERS

| Declaration | Description |
|---|---|
| D1 | The user understands the functionality of the Domain-Specific Language (DSL) elements and their role in application creation process. |
| D2 | This DSL allows to interconnecting devices and people easily, using a few code lines and spending a little time. |
| D3 | Using a DSL makes it difficult to make mistakes while the user is modelling the applications. |
| D4 | This solution offers a fast way to developing the indicated task. |
| D5 | This solution helps create applications to interconnect objects and people. |
| D6 | The DSL does not require the user to use complex programming skills, as in traditional application development. |
| D7 | The DSL includes enough elements and functionality for the user to create a wide range of applications to interconnect objects and people. |
| D8 | This proposal is a positive contribution to encourage the development of services and applications that provide interconnection between objects and people. |
| D9 | Internet of Things will be benefited by this solution. |
| D10 | This DSL could be used to simplify the classic development process of software applications in other areas. |

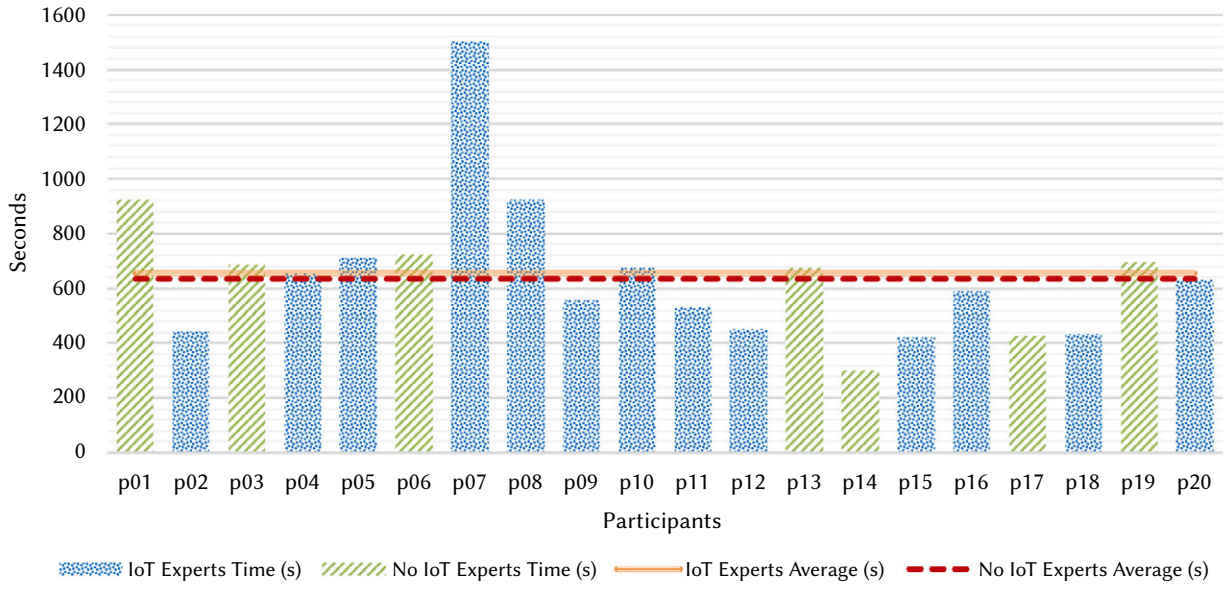



Fig. 1. Time to complete the task per participant.

### B. Results

When the evaluation process finished, results to achieve conclusions was obtained. For that, in the following lines, the results obtained through the evaluation process already defined before will be present. The statistical analysis was performed using R version 3.3.2. These results are going to be analysed by an inter-subject study because of the existence of two different groups, users with knowledge about the IoT and users without this knowledge.

#### 1. Phase 1

Table II shows the times obtained in the first phase, the sample size ($n$), the mean ($\bar{x}$), the standard deviation ($s$), the maximum and minimum for every profile, and all participants. All-time measures represent seconds spent by users to complete the task.

TABLE II. General Descriptive Statistics of Times Spent By Each Profile

| | IoT Experts | No IoT Experts | All participants |
|---|---|---|---|
| $n$ | 13 | 7 | 20 |
| $\bar{x}$ | 656.54 | 634.29 | 648.80 |
| $s$ | 290.80 | 205.70 | 258.60 |
| $max$ | 1504 | 926 | 1504 |
| $min$ | 421 | 303 | 303 |

Moreover, Fig. 1 shows the results of the first phase graphically.

#### 2. Phase 2

The second phase or phase 2 consisted of filling a 5-points Likert Scale survey. As has already been explained earlier, the options of the survey were: Strongly Disagree, Disagree, Neutral, Agree, and Strongly Agree. To make it easier to analyse, numeric values were associated with each option from 1 to 5 according to the worst and the best opinion.

Table III shows the responses of each participant anonymously by indicating the profile of each participant, the numeric value of each answer, and the total score of each participant.

TABLE III. Responses of Participants for Each Declaration

| | IoT Expert | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 | D10 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **p01** | No | 4 | 5 | 5 | 5 | 5 | 4 | 5 | 5 | 5 | 2 | 45 |
| **p02** | Yes | 5 | 5 | 5 | 5 | 5 | 4 | 5 | 5 | 5 | 4 | 48 |
| **p03** | No | 4 | 5 | 4 | 5 | 5 | 3 | 2 | 5 | 4 | 5 | 42 |
| **p04** | Yes | 4 | 4 | 4 | 5 | 3 | 3 | 5 | 3 | 5 | 4 | 40 |
| **p05** | Yes | 5 | 4 | 5 | 5 | 5 | 2 | 4 | 2 | 2 | 1 | 35 |
| **p06** | No | 4 | 5 | 4 | 4 | 4 | 3 | 5 | 4 | 3 | 4 | 40 |
| **p07** | Yes | 4 | 4 | 4 | 4 | 4 | 4 | 3 | 5 | 5 | 4 | 41 |
| **p08** | Yes | 4 | 5 | 3 | 4 | 4 | 5 | 4 | 5 | 5 | 4 | 43 |
| **p09** | Yes | 5 | 5 | 4 | 5 | 5 | 5 | 5 | 5 | 4 | 5 | 48 |
| **p10** | Yes | 4 | 5 | 4 | 5 | 5 | 4 | 5 | 5 | 5 | 5 | 47 |
| **p11** | Yes | 5 | 4 | 2 | 5 | 3 | 3 | 4 | 5 | 5 | 2 | 38 |
| **p12** | Yes | 4 | 5 | 2 | 5 | 5 | 5 | 4 | 5 | 5 | 4 | 44 |
| **p13** | No | 3 | 4 | 4 | 5 | 4 | 5 | 4 | 4 | 4 | 4 | 41 |
| **p14** | No | 5 | 4 | 3 | 5 | 4 | 5 | 5 | 5 | 4 | 5 | 45 |
| **p15** | Yes | 5 | 4 | 4 | 5 | 4 | 4 | 4 | 5 | 5 | 5 | 45 |
| **p16** | Yes | 5 | 5 | 4 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 49 |
| **p17** | No | 4 | 4 | 4 | 4 | 5 | 4 | 4 | 4 | 4 | 4 | 41 |
| **p18** | Yes | 5 | 5 | 4 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 49 |
| **p19** | No | 5 | 4 | 5 | 4 | 5 | 4 | 4 | 4 | 4 | 4 | 43 |
| **p20** | Yes | 5 | 5 | 4 | 5 | 4 | 5 | 4 | 5 | 4 | 4 | 45 |

Fig. 2 shows the distribution of responses for each declaration per user profile. As it shows, most responses are positive although there are some negative opinions.

Table IV shows the descriptive statistics of all declarations and it can be seen the breakdown of each question: the minimum, the first quartile, the median, the third quartile, the maximum, the range (maximum - minimum), the range between quartiles and mode.

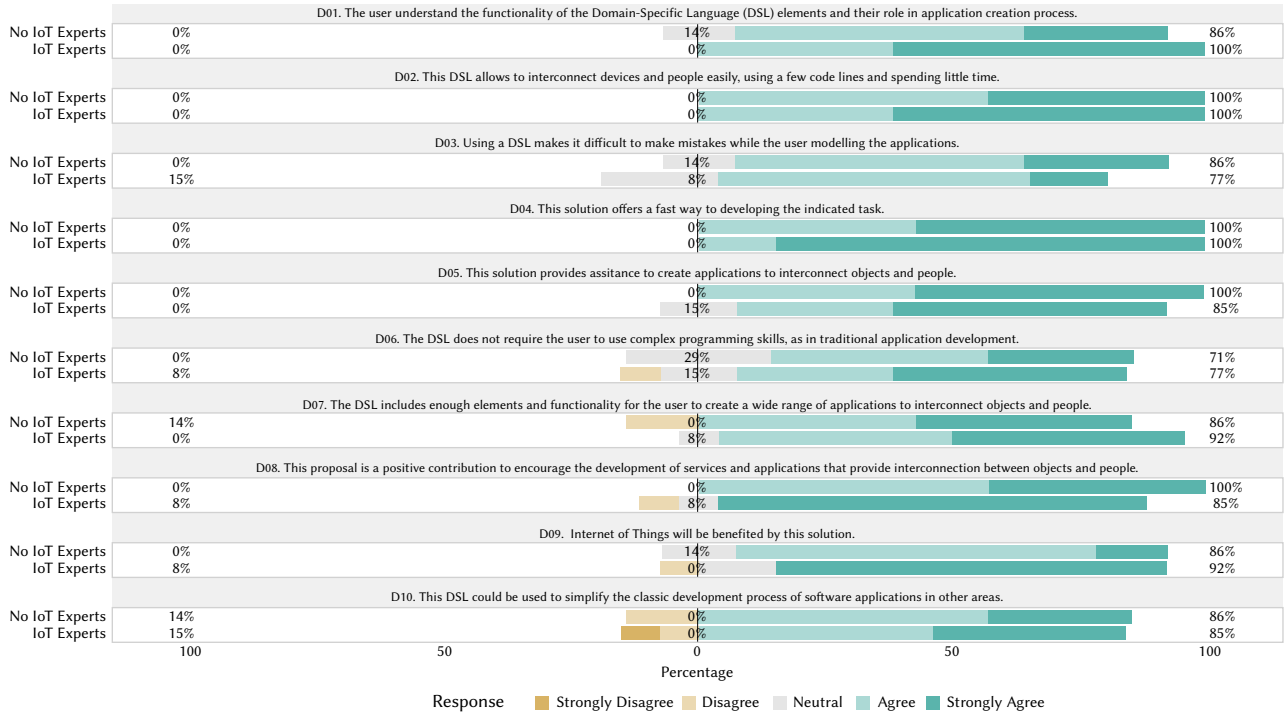



Fig. 2. Distribution of responses for each declaration per user profile.

TABLE IV. General Descriptive Statistics of Each Declaration

| | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 | D10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Min** | 3 | 4 | 2 | 4 | 3 | 2 | 2 | 2 | 2 | 1 |
| **Quartile 1** | 4 | 4 | 4 | 4.75 | 4 | 3.75 | 4 | 4 | 4 | 4 |
| **Median** | 4.5 | 5 | 4 | 5 | 5 | 4 | 4 | 5 | 5 | 4 |
| **Quartile 3** | 5 | 5 | 4 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| **Max** | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| **Range** | 2 | 1 | 3 | 1 | 2 | 3 | 3 | 3 | 3 | 4 |
| **Inter Qrt. - Range** | 1 | 1 | 0 | 0.25 | 1 | 1.25 | 1 | 1 | 1 | 1 |
| **Mode** | 5 | 5 | 4 | 5 | 5 | 5 | 5 | 5 | 5 | 4 |

Fig. 3 shows all this data in a Box and Whiskers Plot diagram.

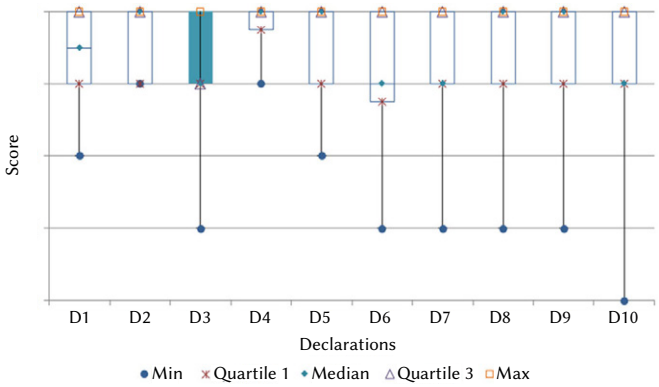



Fig. 3. Box and whiskers plot per declaration.

The frequencies of the responses to each question are shown in Table V. Here, the breakdown of each declaration is shown: the number of votes for each decision and the percentage corresponding to both.

TABLE V. Frequencies TABLE for the General Responses

| | | Strongly disagree | Disagree | Neutral | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| **D1** | # | 0 | 0 | 1 | 9 | 10 |
| | % | 0% | 0% | 5% | 45% | 50% |
| **D2** | # | 0 | 0 | 0 | 9 | 11 |
| | % | 0% | 0% | 0% | 45% | 55% |
| **D3** | # | 0 | 2 | 2 | 12 | 4 |
| | % | 0% | 10% | 10% | 60% | 20% |
| **D4** | # | 0 | 0 | 0 | 5 | 15 |
| | % | 0% | 0% | 0% | 25% | 75% |
| **D5** | # | 0 | 0 | 2 | 7 | 11 |
| | % | 0% | 0% | 10% | 35% | 55% |
| **D6** | # | 0 | 1 | 4 | 7 | 8 |
| | % | 0% | 5% | 20% | 35% | 40% |
| **D7** | # | 0 | 1 | 1 | 9 | 9 |
| | % | 0% | 5% | 5% | 45% | 45% |
| **D8** | # | 0 | 1 | 1 | 4 | 14 |
| | % | 0% | 5% | 5% | 20% | 70% |
| **D9** | # | 0 | 1 | 1 | 7 | 11 |
| | % | 0% | 5% | 5% | 35% | 55% |
| **D10** | # | 1 | 2 | 0 | 10 | 7 |
| | % | 5% | 10% | 0% | 50% | 35% |

Finally, Fig. 4 shows a bar graph with the frequency of the responses in the set formed by both profiles.

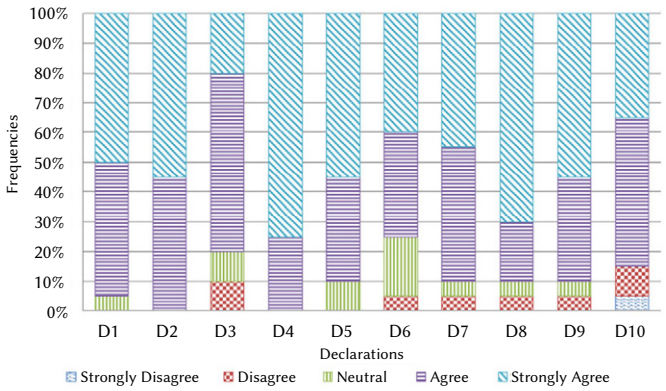



Fig. 4. Frequencies of responses per declaration.

## C. Discussion

This subsection includes a discussion to achieve conclusions.

The aim of the data collected from Phase 1 (Table II and Fig. 1) is to conclude if knowing the IoT affects the time spent by users to complete the task. Before performing a statistical test that verifies that hypothesis, it must be determined if the sample data follow a normal distribution and perform a homoscedasticity test.

The application of the Shapiro-Wilk test shows that the data from users with knowledge about the IoT do not follow a normal distribution ($p = 0.001$) whereas the data from the other profile of users follow, indeed a normal distribution ($p = 0.4$).

As one sample does not follow a normal distribution, the next test used was the Levene test to test the homoscedasticity. The result of that test was the homogeneity of variances ($p = 0.7$).

Finally, to conclude if the relationship between having knowledge about the IoT and the time needed to complete the task is significant, the test used was the Mann-Whitney U test which result was that there are no significant differences ($p = 0.5$) between the time spent by users with knowledge about the IoT and the time spent by users without this knowledge.

Fig. 5 shows a non-significant difference in the time spent to complete the task by the two profiles. There are two outliers that represent the fastest participant and the slowest one.

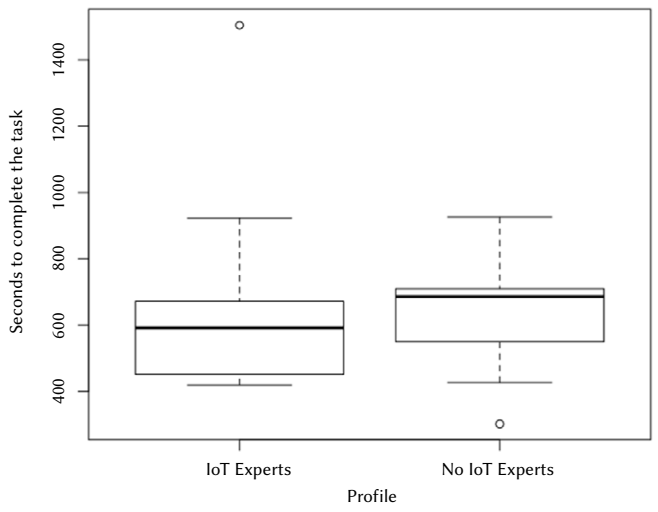



Fig. 5. Box and whiskers plot for times spent by users per profile.

By comparing Fig. 1 and Fig. 5, it can be deduced that the two outliers are the participant p07 and the participant p14. The evaluation process of participant p07 was more laborious than others because he found it hard to understand how the system works. However, participant p14 is a user expert in the development of applications capable of learning a programming language without so much effort. Thus, it was expected that p14 spent less time than other participants to complete the task and it is reasonable to deduce that p07 needed more time than other participants to complete the task.

To conclude the phase 1 of the evaluation process, it can be assumed that having knowledge about the IoT does not affect the use of this proposal because there are no significant differences between both groups. Thus, this proposal is useful for any user without taking into consideration its knowledge about the IoT. However, it is important to remember that this proposal requires basic programming knowledge.

Conclude if the relationship between knowing the IoT and the opinions, expressed via the Likert survey, is significant, is addressed by the collected data in Phase 2.

Nevertheless, it is important to mention that the total score, shown in Table III, is an ordinal variable instead of a cardinal variable because different answers in a Likert Scale do not represent different grades of opinions in an equidistant way, hence, a participant with more score has a better opinion can be assumed but it cannot be said much better is that opinion. Thus, perform statistical analyses that compare the averages to validate the hypothesis is not possible, consequently, a non-parametric test will be used even though the data would fit a normal distribution.

To choose the proper test to perform the statistical analysis, it is necessary to check if the data follow a normal distribution and perform a homoscedasticity test.

The test used to check if the sample data follows a normal distribution was the Shapiro-Wilk test which results was that the data from users with knowledge about the IoT follow a normal distribution ($p = 0.3$) and the data from the other profile of users follow a normal distribution ($p = 0.3$) as well.

As both samples follow a normal distribution, the F test to check the homoscedasticity can be used. The result of this test was the obtained homogeneity of variances ($p = 0.06$).

Finally, to conclude if the relationship between having knowledge about the IoT and the opinions, expressed via the Likert survey, is significant, the Mann-Whitney U test was used because of the homogeneity of variances and the ordinal nature of the values. It was obtained that there are no significant differences ($p = 0.3$) between the opinions of users with knowledge about the IoT and the opinions of users without this knowledge.

Moreover, Fig. 6 also shows that there are no significant differences in the total score obtained by the two user profiles in the Likert Survey.

Now, it can be assumed that having knowledge about the IoT does not influence the users' opinions given via the Likert survey. Thus, the results of the Liker survey can be analysed globally, without discriminating both profiles.

On the other side, from the descriptive statistics of all declaration shown in Table IV and Fig. 3, the following interpretations can be suggested:

- D2 and D4 are the declarations with the highest minimum, in this case, 4 out of 5. This means that all participants agreed with the declaration, at the very least.
- D4 is the declaration with the best score. The first quartile is very close to the maximum value, so the majority chose the maximum option.

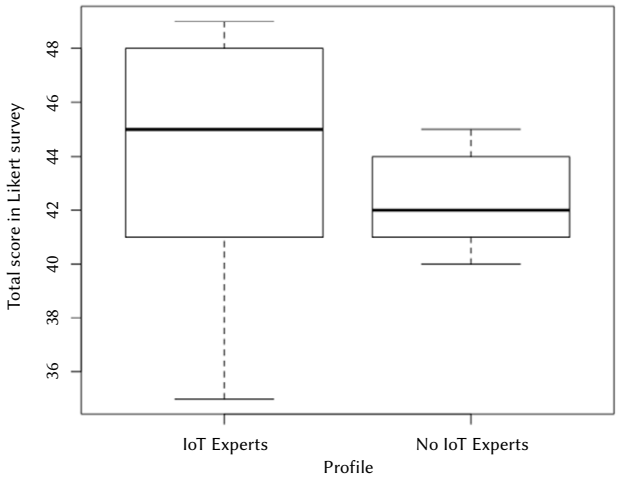



Fig. 6. Box and whiskers plot for total score in Likert survey per profile.

- All questions have a maximum of 5. There is at least one participant that completely agrees with each question.
- D2, D4, D5, D8, and D9 have the highest median, 5 out of 5. From this, it can be deduced that most of the participants agreed with these declarations.
- D2 and D4 have a range of 1 so all participants had the same opinion on these declarations. However, D10 is the only question with a range of 4 which is the worst possible range. It means that there are a lot of differences between the answers of each participant.
- D6 is the question with the biggest dispersion. Only this answer has the first quartile below the answer Agree. Thus, more people chose options different to *Strongly agree* or *Agree* than in the other questions.
- The answer chosen more times is *Strongly agree* because it is the mode of D1, D2, D4, D5, D6, D8 and D9, the mode of D3 and D10 is the answer *Agree* and the mode of D7 are both answers, *Strongly agree* and *Agree*.

Finally, from data shows in Table V and Fig. 4, which contains the frequencies of the responses, it can be figured things that could not be figured before. These are the interpretations:

- D2 and D4 have a 100% of votes for *Agree* and *Strongly agree* and D1 and D5 have 90% or more of votes for *Agree* and *Strongly agree* while only the 10% or less voted *Neutral*. It means that most of the participants agree with theses declarations.
- D7, D8, and D9 have a 10% of votes for *Disagree* and *Neutral*, D3 has 10% of votes for *Disagree* and a 10% of votes for *Neutral*, D6 has less than 10% of votes for *Disagree* but it has a 20% of votes for *Neutral*, and D10 has a 15% of votes for *Disagree* and *Strongly disagree*. It means that the majority agree with theses declarations but there are a few that are indecisive or do not believe in these declarations. Moreover, D10 is the unique declaration with some votes for *Strongly disagree* although the majority votes for *Agree* and *Strongly agree*.

In summary, from the results in Phase 1 we conclude that there are no significant differences between both profiles, and from the results in Phase 2, we conclude a similar result, there are no significant differences between the opinions from both profiles. Users supported this proposal because 80% of the declarations from the survey obtained more than 80% positive or very positive assessments. Moreover, the lowest-rated declaration obtained 75% of positive or very positive assessments.

## IV. State of the Art

### A. Internet of Things

During the last years, one of the most important topics in research and business is the interconnection between heterogeneous and ubiquitous objects between themselves. This technology is better known as the Internet of Things [3]–[5]. The United States National Intelligence Agency [6] has considered the IoT between one of the six technologies with more interest to the United States from here to 2025.

The aim of the IoT is to interconnect heterogeneous and ubiquitous objects and different systems between each other. To achieve that interconnection, it is required things with Internet capabilities [7]. Thus, it can be considered that the IoT was introduced in order to extend the Internet to things [8]. However, not only the interconnection between objects, but also called Machine-To-Machine (M2M) [9]–[11], is important in the IoT, but also the connections between humans and machines (H2M) [12] and among humans (H2H) [10] are also very important because the three types of communications together allow sharing information between the physical world and the virtual world [13]. This novel proposal tries to offer a way of establishing communications amongst objects, and between objects and humans, by using Social Networks directly instead of a common approach of using web services.

However, there is no single standard or way to do that. There are available different Internet platforms that enable the interconnection amongst objects as can be seen in [3], [14]. Moreover, there are many standards to communicate objects like Near Field Communications (NFC), Radio Frequency Identification (RFID), or Bluetooth. For that, it is necessary to facilitate the development in the IoT.

### B. Smart Objects

Smart Objects, also known as Intelligent Products [15], are physical elements that can interact with the environment and/or other objects, have automatic or semi-automatic behaviour depending on the data that they process or receive, and can react according to the interactions with other Smart Objects [4], [16]. Some examples are Smart TVs, smartphones, tablets, some cars, and many other types of devices.

Smart Object can be classified in three dimensions [1] which represent qualities of the object's intelligence: **Level of Intelligence**, **Location of Intelligence**, and **Aggregation level of Intelligence**. The first one indicates how much intelligence an object can have. The second one describes where the intelligence is located. It can be located in the object, in the network, or both. The last dimension indicates if the intelligence is in the element, for example, when the object is composed of various elements and each one has their own intelligence, if it is in the container, or if it is distributed between the container and the elements.

Apart from Smart Objects Not-Smart Objects or objects without intelligence also exist [1]. This type of objects is usually the objects that compose Smart Objects. The Not-Smart Objects are devices that need another device to work like sensors and actuators. Sensors are able to measure physical parameters like the pressure or the temperature, but they cannot process it without another device that is programmed to process data. By the other side, actuators are able to perform actions like control motors or turn on/off lights, but they need another device that orders them the actions to do according to certain conditions.

The proposal of this paper offers Smart Objects whose intelligence is in the container. The other two characteristics depend on the implementation that users developed.

### C. Online Social Networks

Online Social Networks (OSN) are valuable resources to develop applications that could be integrated into people's lives. OSNs provide many services that are useful to create applications like identity and authorisation services, APIs to read or write in timelines, receive updates, receive and send private messages and, so on. OSN is a basic piece of Web 2.0 and the convergence of the real world with OSNs enables the development of new applications capable of interconnecting things and humans [17]. Social Networks are commonly used to gather data about people or events for research purposes. For instance, Twitter can be used to extract information about traffic events using Natural Language Processing [18].

Nowadays, there are many OSNs that could be used for researching but research is focused principally on Twitter followed by Facebook. In this proposal, the OSN chosen was Twitter due to its features. Twitter is an OSN and microblogging service based on short messages of up 140 characters very used to research purposes due to several features. Amongst these features are the next: its philosophy of short public messages, it has a specialised markup language that adds semantic information to messages and makes easier the process of the messages, it has a real-time nature, and the relationships no need reciprocation [19]. Users can follow other users without these users follow back. Due to all these features, Twitter is suitable for this proposal, even though it has some limitations.

As stated above, Twitter is very used in research and even in the frame of the IoT. For example, humans can be considered as a type of sensors that can be useful in Smart Cities [20], to detect Earthquakes [21], or also to support smart decisions about the destination of tourism according to the opinions of Twitter's users [22].

The Social Internet of Things (SIoT) is an approach similar to Online Social Networks but focused on objects instead of humans. These social objects are a new generation of objects that can interact with other objects without the intervention of their owners although with their permission. They are capable of discovering other objects, services, and useful information. Moreover, they also can share their services with the rest of the objects in the network [23]. Based on these principles, in [23], they built their own Social Network for Smart Objects. The disadvantages of this SIoT are the dependence on a specific Social Network which is not used for other purpose, and the interaction of this SIoT is amongst objects whereas in this paper is proposed the intercommunication between humans and objects.

Furthermore, scientists of Ericsson [23] observed that people can familiarise better with IoT technologies if there is an analogy between IoT technologies and their habits in OSNs like Facebook, Twitter or any other.

The combination of OSN and SIoT will bring new interesting applications and possibilities for the IoT.

### D. Related Work

There are not many similar studies that address the integration of devices on traditional Social Networks. However, some investigations address the communication amongst objects [23], [24] and between people and objects [25].

SenseQ [26], [27] is another approach that interconnects people and objects, and uses Twitter as an interface through which users could make queries using natural language that a Wireless Sensor Network (WSN) would try to resolve by collecting data from interconnected and distributed sensors.

A related work is the Midgar IoT platform [3], [14], [28], [29]. Midgar enables the interconnection between heterogeneous and ubiquitous objects with themselves. It uses a graphic DSL to make easier the creation of this interconnection for people without development knowledge. When users have defined their application, Midgar generates a daemon according to users' definition whose aim

is to monitor the database for changes that indicate how to connect the objects. Moreover, Midgar also includes a graphic DSL to create Smart Objects. However, Midgar needs to use a physical server to work and only enables the communications amongst objects whereas Bilrost does not need any server because of the use of Social Networks to communicate objects and users.

Another platform that enables the generation of interconnected Smart Objects is ELIoT [30]. This research presents a novel programming platform for Internet-connected smart devices that are created using a custom language that is based on Erlang. Like the proposal of this paper, this research provides a new language that makes easier its aim. However, even though this proposal provides more features, the proposal of this paper is focused on the generation of Smart Objects that are connected to Social Networks.

A similar approach is [31]. In this article, authors have created a graphic DSL to allow end-users to define rules for Smart Objects. End-users can define the event or events (using the operators 'and', and 'or') to create the rules composed by the events and its linked actions.

There are many other IoT platforms that allow connecting Smart Objects or things to the Internet. In [32], the authors surveyed about 39 IoT platforms. Nevertheless, they did not mention any platform that allows the generation of Smart Objects that can be handled through Twitter.

An approach similar to this work is Social Access Controller (SAC) [25]. It uses Social Networks to share Smart Objects and enables their management. One advantage of SAC is that it enables not only handling Smart Objects remotely but also sharing their status. However, the use of Social Networks is very different. In [25], Social Networks are used to know the friends of owners of the devices and then, it allows friends to access the Smart Objects through REST architecture. However, the proposal of this paper is different because the access to Smart Objects is through Twitter and users that can handle the Smart Objects do not need to friend anyone, they only need to be mentioned in the application definition written with the DSL.

Another point of view is to use **instant messages** as a way of establishing communication between objects with other objects or humans [33]–[35]. This approach has some disadvantages like the dependency of specific applications which are exclusive for this goal whereas the use of Twitter enables the use of the common Twitter application that it is usually available in so many smartphones and it also enables the use of the web application which prevents from depending on a specific technology.

To summarise, Bilrost is a novel approach that enables communications between humans and objects without many requirements by the humans' side. Humans only need access to Social Networks like Twitter. Moreover, this communication is easy for users that use frequently Social Networks because it is based on the common use of Social Networks, post messages in a timeline.

## V. Conclusions

At this stage, the proposal of this paper has already been introduced, Bilrost, a novel one that provides a solution to integrate heterogeneous and ubiquitous Smart Objects into traditional Social Networks. Bilrost enables objects to wait for tweets from users and perform actions according to these tweets.

To achieve an easy integration of objects in Social Networks to people without complex programming knowledge, a new DSL was created, the BSL, which users can use to define their devices with their actuators, the actions that they can perform, and some other properties required to establish the communication with the Social Networks.

Bilrost can generate a project application where the integration into Social Networks is already developed but the specific code for each actuator is not still implemented. Therefore, Bilrost achieves expanding the target audience but it requires that users have a little programming knowledge to fill a skeleton about actuators' actions with a few code lines.

The proposal was evaluated through a two phases evaluation with a sample divided into two different groups with different profiles: IoT experts and no IoT experts. The first phase consisted of performing a task whereas the time spent by the users was being measured and the second one was a Likert survey.

Finally, in the first phase, it was obtained that there are no significant differences between both profiles, and in the second phase it was obtained similar results, there are no significant differences between the opinions from both profiles. Thus, it can be concluded that this proposal is useful for both profiles and their opinions can be analysed altogether. Moreover, in the survey, users supported the proposal because 80% of the declarations obtained more than 80% positive or very positive assessments, and the lowest-rated declarations obtained 75% of positive or very positive assessments. Thus, it can be concluded that Bilrost facilitates the creation of applications that allow non-expert users to control Smart Objects' actuators through Social Networks designed for humans.

Bilrost may be a small step to achieve IoT to be more present in people's day-to-day lives.

## VI. Future Work

The Internet of Things is the future so make the integration of IoT technologies in people's diary lives easier is necessary. This proposal follows this way, but it is still not finished, much future work to do from here remains. In the next items, there is some possible future work that arises from this proposal:

- Upgrade the BSL syntax to enable the definition of sensors. This upgrade will increase the possibilities of generated applications.
- Improve the BSL syntax to enable users to define the specific implementation and improve the Applications Generator to generate end-user applications.
- Create a graphic DSL to make easier the generation of applications.
- Compare and study current Social Networks and choose those that are useful to interconnect people and objects.

Analyse different options to secure the com munications through Social Networks because at this stage these communications are public.

## References

[1] C. González García, D. Meana-Llorián, B. C. P. G-Bustelo, and J. M. C. Lovelle, "A review about Smart Objects, Sensors, and Actuators," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 3, pp. 7–10, 2017, doi: 10.9781/ijimai.2017.431.

[2] D. Meana-Llorián, C. González García, J. Pascual Espada, V. B. Semwal, and M. Khari, "Bilrost: Connecting the Internet of Things through human Social Networks with a Domain-Specific Language," in *Proceedings of the Second International Conference on Research in Intelligent and Computer in Engineering*, 2017, pp. 57–61, doi: 10.15439/2017R110.

[3] C. González García, C. P. García-Bustelo, J. P. Espada, and G. Cueva-Fernandez, "Midgar: Generation of heterogeneous objects interconnecting applications. A Domain Specific Language proposal for Internet of Things scenarios," *Computer Networks*, vol. 64, no. C, pp. 143–158, Feb. 2014, doi: 10.1016/j.comnet.2014.02.010.

[4] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey," *Computer Networks*, vol. 54, no. 15, pp. 2787–2805, 2010, doi: 10.1016/j.

comnet.2010.05.010.
[5] K. Gama, L. Touseau, and D. Donsez, "Combining heterogeneous service technologies for building an Internet of Things middleware," *Computer Communications*, vol. 35, no. 4, pp. 405–417, Feb. 2012, doi: 10.1016/j.comcom.2011.11.003.
[6] National Intelligence Council, "Discruptive civil tehnologies - six tehnologies with potential impacts on us interests out to 2025," *Conference Report CR 2008 - 07*. Apr, 2008.
[7] G. M. Lee and J. Y. Kim, "Ubiquitous networking application: Energy saving using smart objects in a home," in *2012 International Conference on ICT Convergence (ICTC)*, Oct. 2012, pp. 299–300, doi: 10.1109/ICTC.2012.6386844.
[8] S. Li, L. Da Xu, and S. Zhao, "The internet of things: a survey," *Information Systems Frontiers*, vol. 17, no. April 2014, pp. 243–259, 2014, doi: 10.1007/s10796-014-9492-7.
[9] E. Borgia, "The Internet of Things vision: Key features, applications and open issues," *Computer Communications*, vol. 54, pp. 1–31, 2014, doi: 10.1016/j.comcom.2014.09.008.
[10] L. Tan, "Future internet: The Internet of Things," in *2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE)*, Aug. 2010, pp. V5-376-V5-380, doi: 10.1109/ICACTE.2010.5579543.
[11] M. Hasan, E. Hossain, and D. Niyato, "Random access for machine-to-machine communication in LTE-advanced networks: issues and approaches," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 86–93, Jun. 2013, doi: 10.1109/MCOM.2013.6525600.
[12] International Telecommunication Union, "Overview of the Internet of things," Geneva, p. 14, 2012.
[13] I. Mashal, O. Alsaryrah, T.-Y. Chung, C.-Z. Yang, W.-H. Kuo, and D. P. Agrawal, "Choices for interaction with things on Internet and underlying issues," *Ad Hoc Networks*, vol. 28, pp. 68–90, May 2015, doi: 10.1016/j.adhoc.2014.12.006.
[14] C. González García, J. P. Espada, E. R. N. Valdez, and V. García-Díaz, "Midgar: Domain-Specific Language to Generate Smart Objects for an Internet of Things Platform," in *2014 Eighth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, Jul. 2014, pp. 352–357, doi: 10.1109/IMIS.2014.48.
[15] G. G. Meyer, K. Främling, and J. Holmström, "Intelligent Products: A survey," *Computers in Industry*, vol. 60, no. 3, pp. 137–148, Apr. 2009, doi: 10.1016/j.compind.2008.12.005.
[16] C. Y. Wong, D. McFarlane, A. Ahmad Zaharudin, and V. Agarwal, "The intelligent product driven supply chain," in *IEEE International Conference on Systems, Man and Cybernetics*, 2002, vol. vol.4, p. 6, doi: 10.1109/ICSMC.2002.1173319.
[17] M. Blackstock, R. Lea, and A. Friday, "Uniting online social networks with places and things," in *Proceedings of the Second International Workshop on Web of Things - WoT '11*, 2011, p. 1, doi: 10.1145/1993966.1993974.
[18] P. Anantharam, P. Barnaghi, K. Thirunarayan, and A. Sheth, "Extracting City Traffic Events from Social Streams," *ACM Transactions on Intelligent Systems and Technology*, vol. 6, no. 4, pp. 1–27, Jul. 2015, doi: 10.1145/2717317.
[19] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?," in *Proceedings of the 19th international conference on World wide web - WWW '10*, 2010, p. 591, doi: 10.1145/1772690.1772751.
[20] D. Doran, S. Gokhale, and A. Dagnino, "Human sensing for smart cities," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2013, pp. 1323–1330.
[21] T. Sakaki, M. Okazaki, and Y. Matsuo, "Tweet analysis for real-time event detection and earthquake reporting system development," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 4, pp. 919–931, 2013, doi: 10.1109/TKDE.2012.29.
[22] A. Cacho *et al.*, "Social Smart Destination: A Platform to Analyze User Generated Content in Smart Tourism Destinations," in *New Advances in Information Systems and Technologies*, Á. Rocha, M. A. Correia, H. Adeli, P. L. Reis, and M. Mendonça Teixeira, Eds. Cham: Springer International Publishing, 2016, pp. 817–826.
[23] L. Atzori, A. Iera, and G. Morabito, "From 'smart objects' to 'social objects': The next evolutionary step of the internet of things," *IEEE Communications Magazine*, vol. 52, no. 1, pp. 97–105, Jan. 2014, doi: 10.1109/MCOM.2014.6710070.
[24] L. Atzori, A. Iera, and G. Morabito, "SIoT: Giving a Social Structure to the Internet of Things," *IEEE Communications Letters*, vol. 15, no. 11, pp. 1193–1195, Nov. 2011, doi: 10.1109/LCOMM.2011.090911.111340.
[25] D. Guinard, M. Fischer, and V. Trifa, "Sharing using social networks in a composable web of things," in *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2010 8th IEEE International Conference on*, 2010, pp. 702–707, [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5470524.
[26] D. Meana-Llorián, C. González García, V. García-Díaz, B. C. P. G-Bustelo, and J. M. C. Lovelle, "SenseQ: Creating relationships between objects to answer questions of humans by using Social Networks," in *Proceedings of the The 3rd Multidisciplinary International Social Networks Conference on SocialInformatics 2016, Data Science 2016 - MISNC, SI, DS 2016*, 2016, pp. 1–5, doi: 10.1145/2955129.2955135.
[27] D. Meana-Llorián, C. González García, B. C. Pelayo G-Bustelo, and N. García-Fernández, "SenseQ: Replying questions of Social Networks users by using a Wireless Sensor Network based on sensor relationships," *Data Science and Pattern Recognition*, vol. 1, no. 1, pp. 1–12, 2017, doi: http://www.ikelab.net/dspr-pdf/vol1-1/dspr-paper1.pdf.
[28] G. Sánchez-Arias, C. González García, and B. C. Pelayo G-Bustelo, "Midgar: Study of communications security among Smart Objects using a platform of heterogeneous devices for the Internet of Things," *Future Generation Computer Systems*, vol. 74, no. September, pp. 444–466, 2017, doi: 10.1016/j.future.2017.01.033.
[29] C. Gonzalez Garcia, L. Zhao, and V. Garcia-Diaz, "A User-Oriented Language for Specifying Interconnections Between Heterogeneous Objects in the Internet of Things," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 3806–3819, Apr. 2019, doi: 10.1109/JIOT.2019.2891545.
[30] A. Sivieri, L. Mottola, and G. Cugola, "Building Internet of Things software with ELIoT," *Computer Communications*, vol. 89, pp. 141–153, 2016, doi: http://dx.doi.org/10.1016/j.comcom.2016.02.004.
[31] G. Desolda, C. Ardito, and M. Matera, "Empowering end users to customize their smart environments: Model, composition paradigms, and domain-specific tools," *ACM Transactions on Computer-Human Interaction*, vol. 24, no. 2, 2017, doi: 10.1145/3057859.
[32] J. Mineraud, O. Mazhelis, X. Su, and S. Tarkoma, "A gap analysis of Internet-of-Things platforms," *Computer Communications*, vol. 89, pp. 5–16, 2016.
[33] J. Choi and C.-W. Yoo, "Connect with Things through Instant Messaging," in *The Internet of Things*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 276–288.
[34] S. Aurell, "Remote Controlling Devices Using Instant Messaging: Building an Intelligent Gateway in Erlang/OTP," in *Proceedings of the 2005 ACM SIGPLAN workshop on Erlang*, 2005, pp. 46–51.
[35] A. Roychowdhury and S. Moyer, "Instant messaging and presence for sip enabled networked appliances," *Publisher unknown*, 2001, doi: 10.1.1.116.6212.

Daniel Meana-Llorián

Daniel Meana-Llorián is a Visiting Professor at the School of Computer Engineering of Oviedo, University of Oviedo (Spain). He is a Graduated Engineering in Computer Systems, MSc in Web Engineering, and PhD in Computer Science. His research interests include Mobile technologies, Web Engineering, the Internet of Things, and exploration of emerging technologies related to the previous ones.

Cristian González García

Cristian González García is an Assistant Professor in the Department of Computer Science, University of Oviedo (Spain). He is a Technical Engineer in Computer Systems, MSc in Web Engineering, and PhD in Computers Science. He has been a visiting PhD candidate in the University of Manchester. Besides, he has been in the University of South Florida as visiting professor. He has also been working in different national and regional projects, as well as in projects with private companies.His research interests include the Internet of Things, Web Engineering, Mobile Devices, Artificial Intelligence, Big Data, and Modelling Software with DSL and MDE .

B. Cristina Pelayo G-Bustelo

B. Cristina Pelayo G-Bustelo is a Lecturer in the Computer Science Department of the University of Oviedo. Ph.D. from the University of Oviedo in Computer Engineering. Her research interests include Object-Oriented technology, Web Engineering, eGovernment, Modelling Software with BPM, DSL and MDA.

Juan Manuel Cueva Lovelle

Juan Manuel Cueva Lovelle is a Mining Engineer from Oviedo Mining Engineers Technical School in 1983 (Oviedo University, Spain). Ph. D. from Madrid Polytechnic University, Spain (1990). From 1985 he is Professor at the Languages and Computers Systems Area in Oviedo University (Spain). ACM and IEEE voting member. His research interests include Object-Oriented technology, Language Processors, Human-Computer Interface, Web Engineering, Modeling Software with BPM, DSL and MDA.

# Optimal Parameter Estimation of Solar PV Panel Based on Hybrid Particle Swarm and Grey Wolf Optimization Algorithms

Hegazy Rezk[1,2]*, Jouda Arfaoui[3], Mohamed R. Gomaa[4,5]

[1] College of Engineering at Wadi Addawaser, Prince Sattam Bin Abdulaziz University, Wadi Addawaser (Saudi Arabia)
[2] Electrical Engineering Dept., Faculty of Engineering, Minia University, Minia (Egypt)
[3] National School of Engineering of Tunis, BP 37, 1002 Tunis, University of Tunis ELMANAR (Tunisia)
[4] Mechanical Department, Benha Faculty of Engineering, Benha University, Benha (Egypt)
[5] Mechanical Department, Faculty of Engineering, Al-Hussein Bin Talal University, Ma'an (Jordan)

Received 20 June 2020 | Accepted 30 October 2020 | Published 14 December 2020



## Abstract

The performance of a solar photovoltaic (PV) panel is examined through determining its internal parameters based on single and double diode models. The environmental conditions such as temperature and the level of radiation also influence the output characteristics of solar panel. In this research work, the parameters of solar PV panel are identified for the first time, as far as the authors know, using hybrid particle swarm optimization (PSO) and grey wolf optimizer (WGO) based on experimental datasets of I-V curves. The main advantage of hybrid PSOGWO is combining the exploitation ability of the PSO with the exploration ability of the GWO. During the optimization process, the main target is minimizing the root mean square error (RMSE) between the original experimental data and the estimated data. Three different solar PV modules are considered to prove the superiority of the proposed strategy. Three different solar PV panels are used during the evaluation of the proposed strategy. A comparison of PSOGWO with other state-of-the-art methods is made. The obtained results confirmed that the least RMSE values are achieved using PSOGWO for all case studies compared with PSO and GWO optimizers. Almost a perfect agreement between the estimated data and experimental data set is achieved by PSOGWO.

## Keywords

Modern Optimization, Parameter Estimation, Renewable Energy, Energy Efficiency, Single-diode Model, Double-diode Model.

DOI: 10.9781/ijimai.2020.12.001

## I. Introduction

THE expanding need for power and the necessity to save the environment have led to an increased focus on renewable energy resources. Solar energy is addressed as a crucial and promising alternative, especially for the electrical power domain, regarding its merits in terms of availability and cleanliness. In this context, the prevailed tendency is to produce some strategies, aimed at ensuring the effectiveness of photovoltaic devices design. The production chain's effectiveness for electricity relies upon the reliability of solar cells (SCs). For obtaining the maximum output energy, it is mandatory to design accurately and with efficacy the photovoltaic (PV) module

Regarding this matter, a prerequisite is to produce a proper mathematical model and reliable patterning techniques enabling the simulation of the actual behavior of photovoltaic cells or modules. The single diode model (SDM) and the double diode model (DDM) have been considered as the widely employed mathematical models [1]. The PV modeling could prove crucial in achieving an appropriate and effective conception for the PV systems. To assess the efficiency of the PV models, the process of extracting the PV parameters becomes a hard task due to the non-linearity aspect of (*I-V*) characteristics. To overcome this issue, the estimation procedure of these variables for both SDM and DDM is addressed as a non-linear optimization problem under different operating constraints, aimed to adjust these decision variables.

PV modeling strategies are classified according to either the available data (manufacturers, experimental measures) or the established method, thus obtaining an accurate PV model. In that regard, a variety of approaches are introduced in the literature aimed to extract optimal PV cells parameters of such complex design, and that provides a significantly improved accuracy. These approaches are commonly categorized into three groups: Analytical, numerical, and hybrid methods [2].

Concerning the traditional methods (analytical methods), the identification of PV parameters required elementary functions [3], taking into consideration some points of both (*I-V*) and (*P-V*) curves, which are identified as important. These methods have the advantage of being easy to be implemented and having a reduced computational cost. An impressive selection of points offers high-quality solutions. The main drawback of the analytical techniques is using few

* Corresponding author.
E-mail address: hr.hussien@psau.edu.sa

assumptions that are made to reduce the number of the unknown parameters. For this purpose, several deterministic and metaheuristics methods are examined, aimed to assess the efficacy of such PV devices.

The iterative approaches such as newton-raphson with likehood estimator [4], gauss-seidel [5] approaches were applied to solve the restrictions of analytical approaches. Deterministic methods are designed to estimate the parameters which govern the PV model, and they consist of the Levenberg-Marquardt method (LM) [6], Newton-Raphson method [7] and Conductivity Method (CM) [8]. Furthermore, the solutions obtained via these approaches are in heavy dependence on the initial conditions of the unknown parameters and easily catch the local optimal solution. Such methods are not appropriate for parameter extraction of PV models under any environmental conditions. In deterministic method [9], the parameters have been estimated by considering a large number of actual measured data. These methods give a faster response, however, the results accuracy of these methods are yet to improve since deterministic methods follow gradient-based algorithms which will easily dive into the local optimal solution, having more restrictions. Additionally, the precision of these approaches is less as the initial solutions are far from the global best solutions [10].

For more accurate and reliable solutions, soft computing methods are introduced for this purpose, which are based on global optimization theory. Based on the literature, there are several metaheuristics algorithms which can be categorized into four groups: Evolutionary-based algorithms; Swarm intelligence-based algorithms; Physics-based algorithms and Human-based algorithms. Researches highlighted the significance of these methods in parameter extraction of PV SCs such as Genetic algorithm (GA) [11], Artificial Bee Colony (ABC) [12] and other optimizers [13]-[19].

Some strategies suffer from shortcomings regarding: quality of solution (catching the local optimum) and convergence speed; computational execution time (execution time is often longer); performance under different environmental conditions. Recently, hybrid techniques were prevailed in addressing the PV parameter extraction issue. Distinct strategies or different optimization techniques are incorporated, thus forming a hybrid method. Various researchers are focused on the use of hybrid techniques to manage the limitations of the previous methods. For example, in [20], the authors introduced Levenberg–Marquardt algorithm combined with simulated annealing (LMSA) hybrid strategy, which is the result from a combination of LM and SA metaheuristic techniques. The hybrid strategy (EHA-NMS) [21] is based on the combination of the two swarm methods Eagle Strategy (ES), ABC and deterministic NMS (Nelder-Mead simplex) technique.

Authors in [26] proposed a novel hybrid approach based on the Pattern Search method and the Firefly algorithm to extract the parameters of both SDM and DDM. To validate the effectiveness, this new approach was compared with other optimizations algorithms used for the parameters extraction process. The proposed strategy outperforms the considered studies techniques in terms of quality solution and accuracy.

In that context, the particle swarm optimization method (PSO) is considered as the most widely spread metaheuristic (MH) technique due to its ease of implementation. Despite that, this technique suffers from some shortcomings such as the premature convergence and the catching of the local optimum. To avoid these drawbacks, several PSO variants are proposed such as the GA-PSO hybrid method [27], which is generated by the combination of the GA and PSO algorithms, seeking to identify the parameters of the single diode PV modules, this hybrid technique performs better than the classical GA metaheuristic. Another hybrid technique named Guaranteed Convergence Particle Swarm Optimization (GCPSO) was examined in [28]. The objective of this technique is to estimate the PV parameters of both SDM and DDM, under different functioning conditions. Consequently, the particle swarm stagnation and the premature convergence were evaded. Also, this strategy has exhibited better performance in terms of accuracy and computational time.

In sum, the process of parameters' estimation of PV modules has been proved as a hard challenge by several literature reviews. This problem is reformulated as an optimization problem with constraints, which can be managed effectively thanks to such advanced metaheuristics methods. Many researchers focus on how to invent new methods, which can estimate the unknown parameters of solar cells, with high accuracy as well as non-premature convergence of solutions. In this paper, the parameters of solar PV panel are identified for the first time, as far as the authors know, using hybrid particle swarm optimization (PSO) and grey wolf optimizer (WGO) based on experimental datasets of *I-V* curves. The chief benefit of hybrid PSOGWO is combining the exploitation ability of the PSO with the exploration ability of the GWO. Three different solar PV modules are considered to prove the superiority of the proposed strategy. Three different solar PV panels are used during the evaluation of the proposed strategy. A comparison of PSOGWO with other state-of-the-art methods is made.

## II. Modelling of PV Panel

To achieve an effective design of PV systems, a lot of literature review works are looking to develop mathematical modeling of solar PV modules. The single diode model (SDM) is the most commonly used one due to the ease of implementation as well as the compromise reached between the accuracy and simplicity. However, to enhance the accuracy representation, the double diode model (DDM) appeared and is considered for uses, especially under low irradiation.

### A. Single Diode Model

Fig. 1 depicts the electrical equivalent model of a single diode of PV cells. By applying Kirchhoff's law, this model is expressed by Eq (1):

$$I = I_{pv} - I_D - \left(\frac{V+I.R_s}{R_p}\right) \tag{1}$$

Where

$I_{pv}$ and $I_D$ represent the photo-generated current and the diode current, respectively.

$R_s$ and $R_p$ indicate the series and shunt resistance.

A current supply $I_{pv}$ is linked to a parallel diode $D$ with ($I$-$V$) characteristic curve, which is defined by Shockley in the following formula as:

$$I_D = I_{o1}\left[\exp\left(\frac{(V+I.R_s)}{n_1.V_t}\right) - 1\right] \tag{2}$$

The ideality factor of such diode is denoted by $n_1$, selected according to the sort of semi-conductor material and the fabrication design.

$V_t$ represents the thermal voltage, which expresses as follows [29]:

$$V_t = \frac{N_s.K.T}{q} \tag{3}$$

$K$ is the Boltzmann constant and it is equal to $1.35^{*}10^{-23}$.

$T$ indicates the PV cell temperature, expressed in kelvin.

$N_s$ and $q$ represent the number of PV cell which are connected in series and the charge of electron ($1.6^{*}\ 10^{-19}$).

The model shown in Fig. 1 is characterized by five variables expressed as follows ($I_{pv}$, $I_{o1}$, $n_1$, $R_s$, $R_p$), that can be identified by analytic or numerical method.

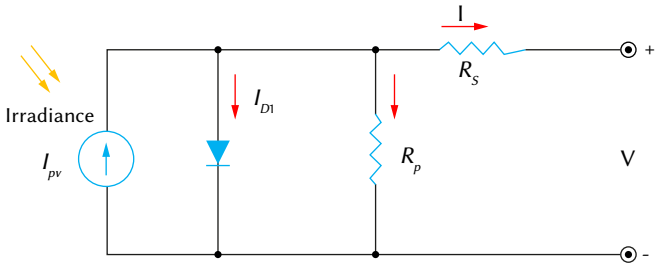



Fig. 1. Equivalent circuit of single diode model (SDM).

### B. Double Diode Model

The electrical equivalent model of a double diode is similar to that of the single diode, adding the fact of having two diodes connected in parallel to the current generator. This sort of model is able to simulate the behavior of PV modules under different irradiation conditions [15]. To achieve more accuracy, DDM is highly useful even if the number of unknown parameters would be increased. Fig. 2 illustrates the equivalent circuit model of the double diode.

Similarly, the generated current is obtained by applying Kirchhoff's law, and it is described as follows:

$$I = I_{pv} - I_{o1}\left[\exp\left(\frac{q(V+I.R_s)}{n_1.k.T}\right) - 1\right] - \ldots\ldots$$
$$I_{o2}\left[\exp\left(\frac{q(V+I.R_s)}{n_2.k.T}\right) - 1\right] - \left(\frac{V+I.R_s}{R_p}\right) \quad (4)$$

Where $n_1$ and $n_2$ represent the ideality factor of diode D1 and diode D2, respectively.

The diffusion and saturation current values are indicated by $I_{o1}$ and $I_{o2}$.

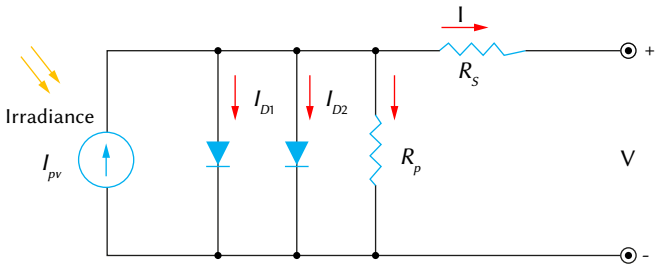



Fig. 2. Equivalent circuit of double diode model (DDM).

To ensure an efficient modeling of DDM, these parameters ($I_{pv}$, $I_{o1}$, $I_{o2}$, $n_1$, $n_2$, $R_s$, $R_p$) shall be determined.

## III. Problem Formulation

To build an accurate PV mathematical model, the (*I-V*) characteristic of PV cells is needed. The non-linear aspect of this equation conducts to a non-linear mathematical model, governed by several unknown variables. Therefore, the estimation process of these parameters $\alpha$= ($I_{pv}$, $I_{o1}$, $n_1$, $R_s$, $R_p$) for the SDM and $\alpha$= ($I_{pv}$, $I_{o1}$, $I_{o2}$, $n_1$, $n_2$, $R_s$, $R_p$) for the DDM, is reformulated as a non-linear optimization problem.

To effectively resolve such a hard optimization problem, several optimization algorithms were investigated.

The performance requirements in terms of accuracy identification should be achieved through an appropriate design of an objective function that shall be minimized. The implementation cost function is described by Eq (6), which adopts the root mean square error criteria.

The difference equation between the detected current $I_{det}$ and the predicted current $I_{pre}$, which can be quantified trough several performance indexes, is defined as follows:

$$J(\alpha) = I_{det} - I_{pre}(V_{det}, \alpha) \quad (5)$$

The appropriate cost function can be provided as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(J_i(\alpha))^2} \quad (6)$$

Where $N$ is the set of empirical detected points ($I_i$, $V_i$).

The predicted current value is obtained by means of Eq. (1) and Eq. (4) of the SDM and DDM, as a measure of the detected voltage ($V_{det}$) and the estimated variables, respectively.

## IV. Hybrid Particle Swarm Optimization and Grey Wolf Optimizer

### A. Standard Particle Swarm Optimization Algorithm

Particle swarm optimization (PSO) is a metaheuristic algorithm, which is initially developed by Kennedy and Eberhart [30]. PSO was motivated by social behavior flocks' birds, which serves as a set of design variables.

The PSO technique uses $n_p$ particles, randomly distributed in the research space initially considered, to find an optimal solution. Each particle, representing a candidate solution, is characterized by a position and velocity.

The next position of the particle $x_i^{t+1}$ is obtained from the current position $x_i^t$ as well as from the new calculated velocity $v_i^{t+1}$.

In fact, the next velocity of each agent $v_i^{t+1}$ is computed as a function of its current velocity $v_i^t$, the current position $x_i^t$, the distance to the best personal particle's performance at iteration t, $pbest_i$ and the distance to the best particle in the particle's neighborhood at iteration t, *gbest*.

$$v_i^{t+1} = \overbrace{w.v_i^t}^{first\_part} + \overbrace{C_1.rand_1.(pbest_i - x_i^t)}^{second\_part} + \ldots\ldots$$
$$\overbrace{C_2.rand_2.(gbest - x_i^t)}^{third\_part} \quad (7)$$

$$x_i^{t+1} = x_i^t + v_i^{t+1} \quad (8)$$

Where $W$ is the inertia factor used to control the influence of particle's velocity on its next move, in order to maintain a balance between the exploration and exploitation of the search space.

$C_1$ and $C_2$ are the cognitive and the social coefficient respectively.

$rand_1$ and $rand_2$ are two random variables distributed according to a uniform distribution law in the interval [0 1].

The first part of Eq (7) provides the exploration capability of the PSO algorithm.

The second part of Eq (7) moves the particle towards the best position ever achieved by himself, and the third part of Eq (7) moves the particle according to the best position achieved by all the particles in the population.

Further, the PSO method is initialized by an initializing population of particles whose velocities are computed using Eq (7). The process update of particles' positions is defined as Eq (8). Finally, PSO will be stopped by achieving an end criterion.

### B. Grey Wolf Optimizer

Motivated by grey wolves, the metaheuristic GWO imitates the hunting process and the leadership hierarchy of grey wolves [31]. Grey wolves exist at the highest level of the food chain and regarded as predators. To make sure that the hunting mechanism performs, greys wolves opt to live within the pack.

The mathematical model of the hunting mechanism of GWO consists of a leader wolf ($\alpha$ group), which represents the best fittest solution and a group of followers ($\beta$, $\delta$ and $\gamma$ groups) that are trying to offer the best location of prey via hunting procedure.

In fact, each hunting mechanism consists of two main components parts: tracking and catching the prey, then encircling and attacking the prey until the stop of its moving act.

Over the hunting process, preys have been encircled by the grey wolves. The following equations developed in [31] simulate the encircling's behavior:

$$D = \left|C * X_p\,(t) - X(t)\right| \tag{9}$$

$$X(t+1) = X_p(t) - A * D \tag{10}$$

Where $t$ presents the current iteration, $X_p$ and $X$ indicate the position of prey and the location of grey wolves, respectively.

$A$ and $C$ indicate the coefficients vectors, which are computed as follows:

$$A = a * (2 * r_1 - 1) \tag{11}$$

$$C = 2 * r_2 \tag{12}$$

$r_1$ and $r_2$ are random numbers, selected within the interval [0 1].

The component "a" shall be decreased in a linear manner starting from 2 to 0, over the different iterations.

To discover the prey's location, alpha wolves seek to lead the grey wolves, the other wolves' groups are needed to ensure that this procedure runs perfectly.

By using this GWO metaheuristic, the best solution is guaranteed by the alpha wolves, beta and delta wolves reported the second and third-best solutions.

The process update of grey wolves' position reported in [35] is presented as:

$$\begin{aligned} D_\alpha &= \left|C_1 * X_\alpha - X(t)\right| \\ D_\beta &= \left|C_2 * X_\beta - X(t)\right| \\ D_\delta &= \left|C_3 * X_\delta - X(t)\right| \end{aligned} \tag{13}$$

For each iteration, the best three wolves are represented by $X_\alpha$, $X_\beta$, and $X_\delta$:

$$\begin{aligned} X_1 &= \left|X_\alpha - a_1 * D_\alpha\right| \\ X_2 &= \left|X_\beta - a_2 * D_\beta\right| \\ X_3 &= \left|X_\delta - a_3 * D_\delta\right| \end{aligned} \tag{14}$$

In fact, the updated position of the prey is provided by the mean of three values of positions assessed as the best solutions, which is defined as follows:

$$X_p(t+1) = \frac{X_1 + X_2 + X_3}{3} \tag{15}$$

Attacking prey is considered as the last stage of the GWO method. The condition that guarantees this process is formulated as follows: enough closing to the prey, when the prey achieves an adequate close for values less than 1, grey wolves found themselves in an attack position of preys. This algorithm has the advantage to avoid the wolves getting catch the local minimum when the GWO approach stopped by achieving an end criterion.

### C. The Hybrid PSOGWO Algorithm

The significantly referred variant of PSO is denoted PSOGWO. The fundamental principle of this hybridization method is to integrate the capability of social thinking ($g_{best}$) for PSO with the local search ability of GWO.

The hybrid PSOGWO method has been examined without making changes in the basics operations of the Standard PSO and GWO techniques. In this context, the PSO algorithm is considered as the most used MH technique due to its simplicity and ease of implementation. However, when the PSO algorithm is subjected to some constraints, this technique suffers from shortcomings such as catching the local minimum. In this regard, to avoid this drawback, GWO is proposed to reduce the chance of trapping on the local minimum. In addition, this technique has the advantage of preserving a balance between the exploitation and exploration mechanisms over the optimizing process. The different steps of the hybrid PSOGWO method are illustrated in Fig. 3. PSO algorithm ensures that particles are directed to random positions with a small chance to prevent the local minimum. These directions may have present risks lead to move closer to the local minimum instead of the global minimum. Due to its exploration ability, the GWO algorithm is considered to avoid these risks by replacing these particles by the other ones having improved positions by the run of the GWO algorithm. Since the GWO technique is still used as a complement to the PSO technique, the time execution is extended. However, the successful results and the additional time required are taken into account, and the extended execution time can be considered as acceptable depending on the nature of the optimization problem that shall be resolved.

## V. Results and Discussion

To prove the validity of the proposed PSOGWO approach, it is applied to determine the parameters of different solar PV equivalent circuit models, including the SDM and DDM. Three different experimental datasets are adopted. For the first case, an experimental standard dataset of a Photowatt-PWP 201. It contains 36 polycrystalline silicon cells and operated at 45˚C and 1000 W/m$^2$ [32]. A four solar cell of STE4/100 is used for the second case study. These data are taken from [33]. The test has been performed at 22 $^\circ C$ under irradiance of 900 W/m$^2$. For the third experimental dataset, these data are extracted using FSM solar PV module at a temperature of 30 $^\circ C$. The number of measured voltage and current points is 21. The experimental test rig is shown in Fig. 4. More details about the monitoring system used for recording the experimental dataset can be found [34]. Table I highlights the different specifications of PV panels that are considered in this research.

### A. Results of 1st Dataset

Based on the experimental dataset on Photowatt-PWP 201 PV module, the proposed strategy of PSOGWO is used to determine the optimal parameters of the cell for both SDM and DDM. Table II shows the maximum and minimum boundaries of each unknown parameter and optimal values of SDM and DDM parameters of Photowatt-PWP 201 PV module using PSOGWO, ALO-LW [32], PS [35] and GA [36].

For SDM, the minimum value of RMSE is achieved by PSOGWO strategy. The values of RMSE are 3.06E-03, 1.43E-02, 1.18E-02, and 6.84E-03 respectively for PSOGWO, ALO-LW, PS, and GA. The coefficient of determination is 0.999952 using PSOGWO. This confirms that there is an almost perfect agreement between the estimated datasets and the experimental data. Fig. 5(a) and (b) show the experimental dataset versus the estimated respectively for both SDM and DDM. For the DDM, the RMSE and MAE are 2.87E-03 and 2.33E-03, respectively.

The absolute error against measured PV module voltage for both SDM and DDM using different strategies is shown in Fig. 6. For SDM, the maximum values for the absolute error are 0.0125, 0.0066, and 0.006 respectively for GA, PS, and PSOGWO. Whereas the maximum absolute error for DDM is 0.0056 using the proposed strategy. This also confirms the superiority of PSOGWO compared with other methods.

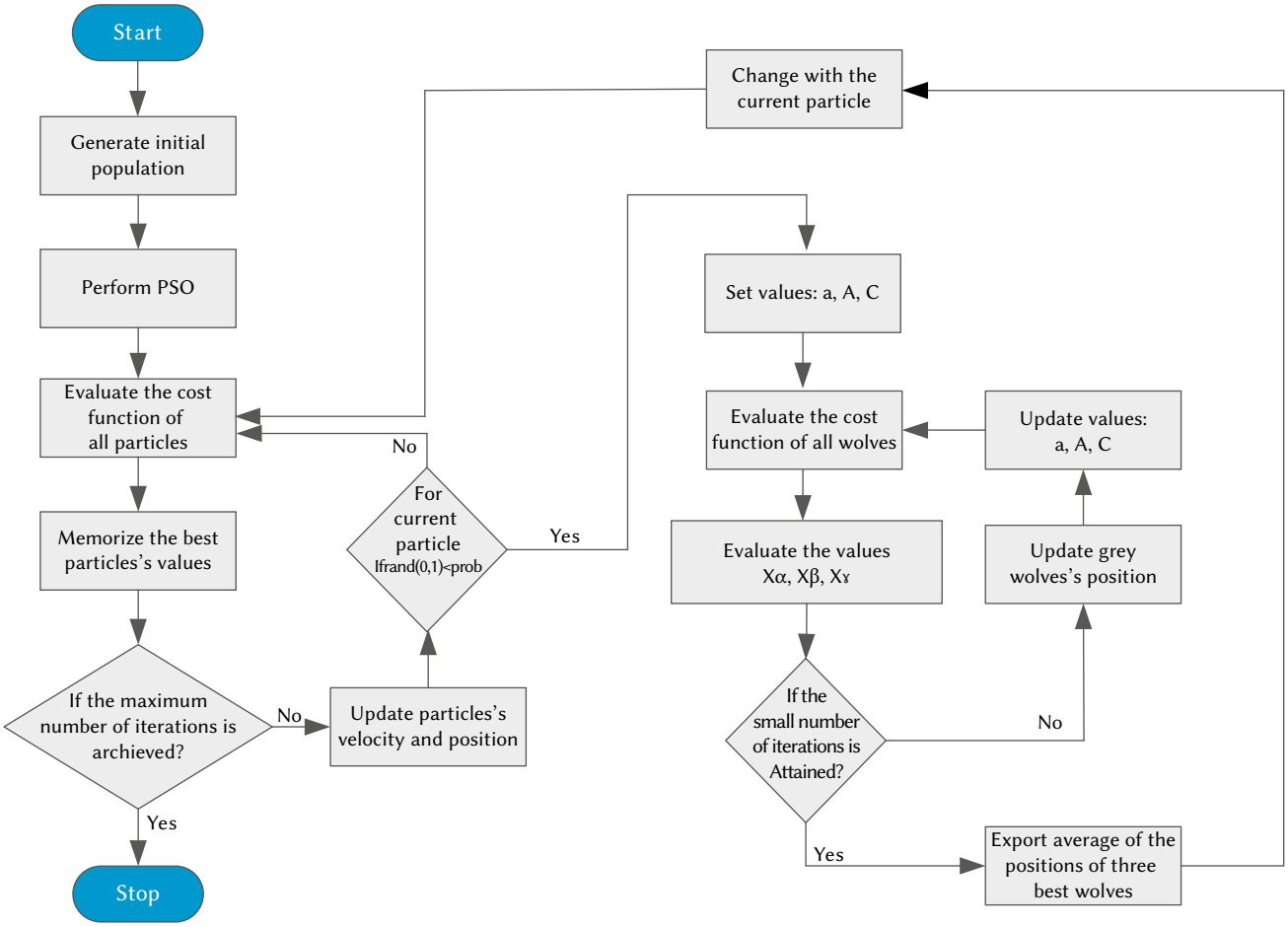



Fig. 3. The steps of PSOGWO.

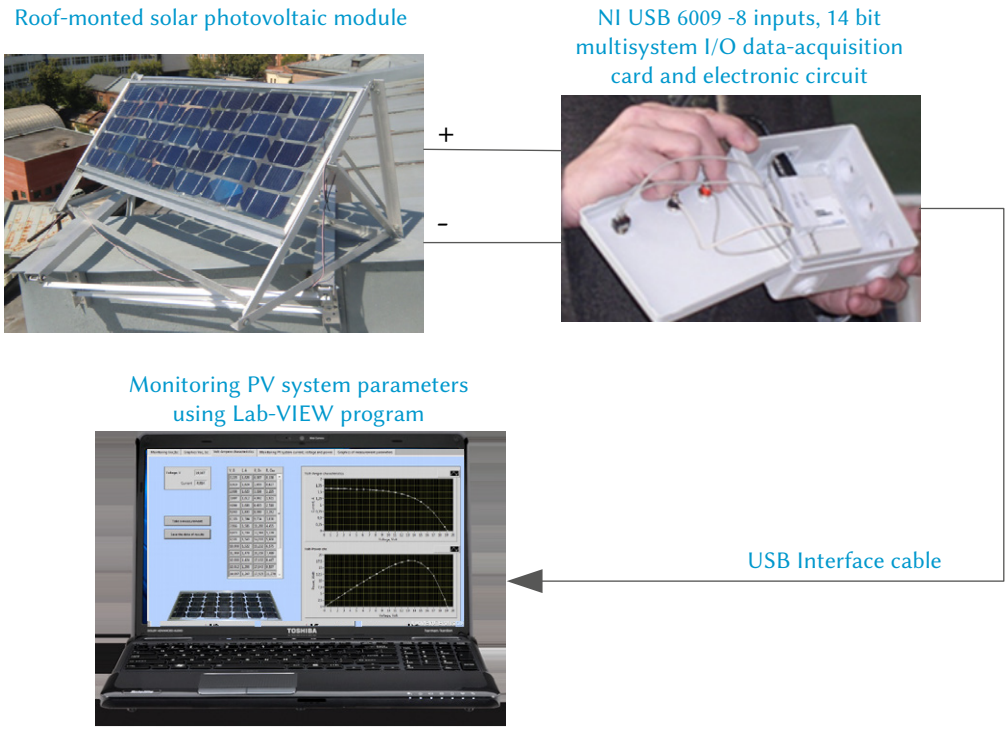
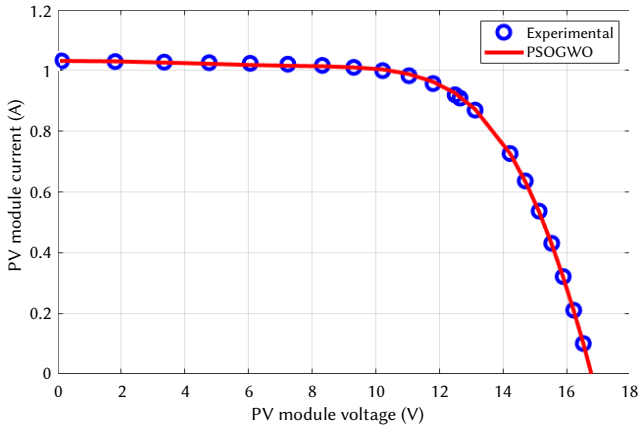
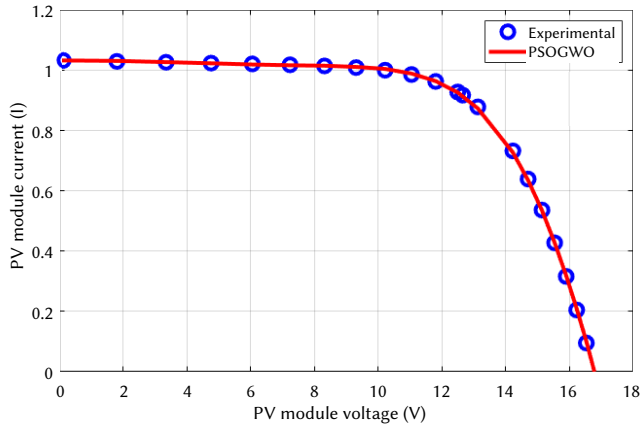



Fig. 4. The experimental test rig.




a) SDM




b) DDM

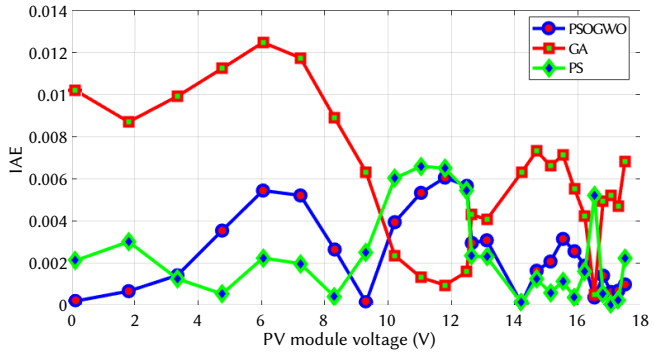
Fig. 5. The experimental dataset versus the estimated for Photowatt-PWP 201 PV module.

TABLE I. Specifications of Photowatt-PWP 201, STE4/100 and FSM

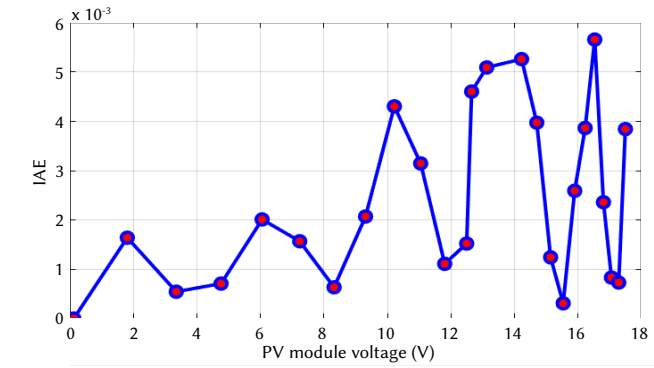
| Parameter | Model of solar PV panel | | |
|---|---|---|---|
| | Photowatt PWP 201 | STE4/100 | FSM |
| **Number of samples** | 26 | 18 | 21 |
| **Test Temperature, C** | 45 | 22 | 30 |
| **Test radiation, W/m2** | 1000 | 900 | na |
| **Short circuit current, A** | 1.0315 | 26.4E-3 | 1.105 |
| **Open circuit voltage, V** | 16.79 | 2.0 | 19.02 |
| **Current @ MPP, A** | 12.4929 | 27.7E-3 | 0.917 |
| **voltage @ MPP, V** | 0.9255 | 1.6 | 14.00 |
| **Number of cells** | 36 | 4 | 35 |

Fig. 7 shows the variation cost function during parameter estimation of Photowatt PWP 201 PV panel using PSOGWO strategy for both SDM and DDM. For both models, approximately 1000 iterations are required to catch the best solution. The best solution values are 3.06E-03 and 2.87E-03 for SDM and DDM, respectively.

The results of the whiteness test for Photowatt PWP 201 PV panel using PSOGWO strategy are shown in Fig. 8. The main target of this test is to ensure that the selected model parameters describe the experimental dataset. It is calculated using the residual autocorrelation function (RACF) at different time lags. Considering Fig. 8, the RCAF values range from -1 to +1.




a) SDM




b) DDM

Fig. 6. Absolute error against measured PV module voltage for both SDM and DDM Photowatt-PWP 201 PV module using different strategies.

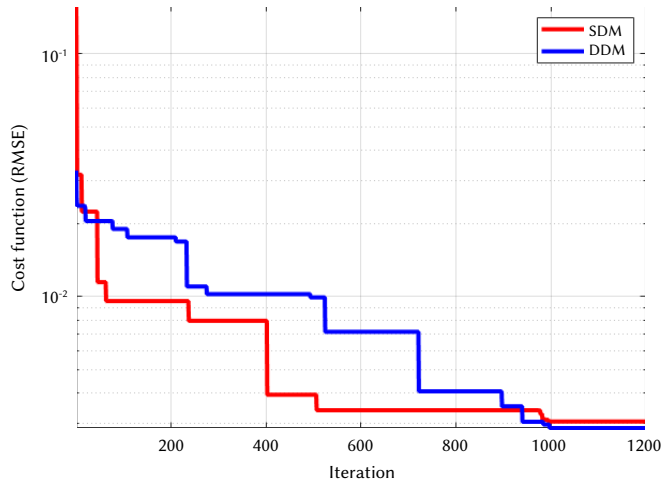
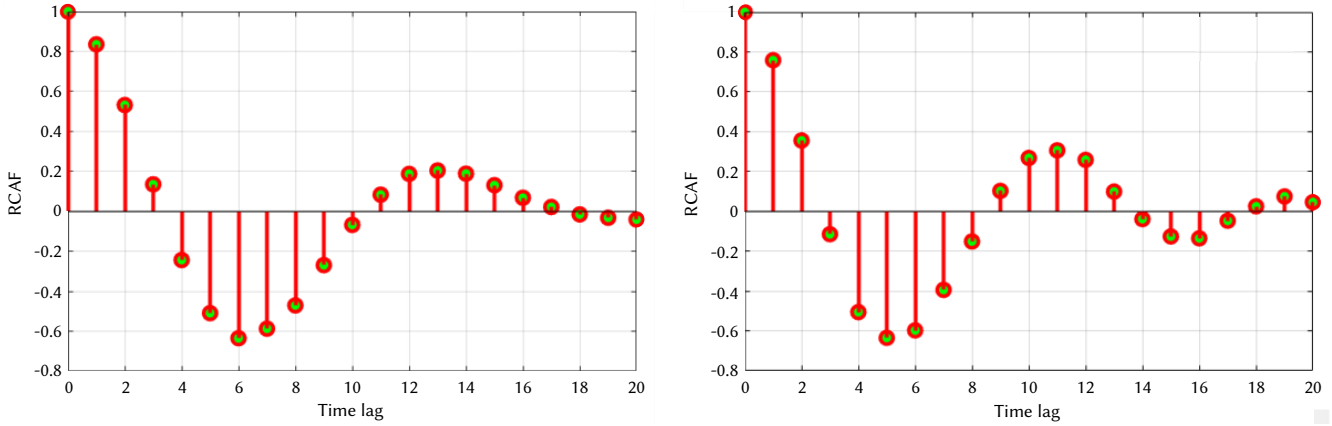



Fig. 7. The variation of cost function during parameter estimation of Photowatt PWP 201 PV panel using PSOGWO strategy.

## B. Results of 2nd Dataset

Based on the experimental dataset on STE4/100 PV solar module PV module, the proposed strategy of PSOGWO is used to determine the optimal parameters of the cell for both SDM and DDM. The number of I-V points is 22. Table III shows the maximum and minimum boundaries of each unknown parameter and optimal values of SDM parameters of STE4/100 PV solar module using PSOGWO, GWO, and ACT [33].

The minimum value of RMSE is achieved by PSOGWO strategy. The values of RMSE are 3.0574E-4, 6.0221E-4, and 3.33925E-4, respectively, for PSOGWO, GWO, and ACT method. The best coefficient of




a) SDM b) DDM

Fig. 8. RCAF results of SDM and DDM for Photowatt-PWP 201 PV module using SFS strategy.

TABLE II. Boundaries and Optimal Values of SDM and DDM Parameters of Photowatt-PWP 201 PV Module

| Parameter | Boundary | | Optimal parameters (SDM) | | | DDM | |
|---|---|---|---|---|---|---|---|
| | Min. | Max. | PSOGWO | ALO-LW [32] | PS [35] | GA[36] | PSOGWO |
| $I_{sc}$ (A) | 0.0 | 1.5 | 1.0328 | 1.03354 | 1.0313 | 1.0441 | 1.0343 |
| $Io_1$ (A) | 1.0e-9 | 1.0e-3 | 5.736e-06 | 4.53123E-6 | 3.1756E-6 | 3.4360E-6 | 9.2835e-07 |
| $Io_2$ (A) | 1.0e-9 | 1.0e-3 | Na | na | na | na | 3.4626e-07 |
| $a_1$ | 0.0 | 3.0 | 1.4074 | 49.8068 | 48.2889 | 48.5862 | 1.2237 |
| $a_2$ | 0.0 | 3.0 | Na | na | na | na | 1.7401 |
| $R_s$ (Ω) | 0.0 | 5.0 | 1.1257 | 1.1246 | 1.2053 | 1.1968 | 1.3398 |
| $R_{sh}$ (Ω) | 0.0 | 2000 | 868.165 | 415.529 | 714.286 | 55.55 | 535.667 |
| RMSE | | | 3.06E-03 | 1.43E-02 | 1.18E-02 | 6.84E-03 | 2.87E-03 |
| MAE | | | 2.42E-03 | 1.69E-01 | 2.27E-03 | 6.14E-03 | 2.33E-03 |
| $R^2$ | | | 0.999952 | na | na | na | 0.999957 |

TABLE III. Boundaries and Optimal Values of SDM Parameters of STE4/100 PV Solar Module

| | Boundary | | Optimal parameters | | |
|---|---|---|---|---|---|
| parameter | Min. | Max. | PSOGWO | GWO | ACT [33] |
| $I_{sc}$ (A) | 0.0 | 1.0 | 26.419E-3 | 26.1449E-3 | 0.024.64E-3 |
| $Io_1$ (A) | 1.0E-10 | 1.0E-5 | 9.4392E-09 | 1.00e-08 | 1.29814E-8 |
| $a_1$ | 0.0 | 2.0 | 1.3369 | 1.58663 | 1.0304 |
| $R_s$ (Ω) | 0.0 | 5.0 | 0.7322 | 0.001 | 2.5568 |
| $R_{sh}$ (Ω) | 0.0 | 5000 | 2488.087 | 4900 | 2184.82 |
| RMSE | | | 3.0574E-4 | 6.0221E-4 | 3.33925E-4 |
| MAE | | | 2.13123E-04 | 3.9455E-4 | 1.98027E-4 |
| $R^2$ | | | 0.99828 | 0.993328 | 0.99800 |

determination is 0.99828. It is achieved by using PSOGWO. This confirms that there is an almost perfect agreement between the estimated datasets and the experimental data. Fig. 9 shows the experimental dataset versus the estimated respectively for SDM.

The absolute error against measured PV module voltage for SDM using different strategies is shown in Fig. 9. As shown in Fig. 10, the maximum values for the absolute error are 7.15E-04, 1.50E-03, and 1.00E-03, respectively for PSOGWO, GWO, and ACT method. This confirms the superiority of PSOGWO compared with GWO and ACT method. The variation cost function during parameter estimation of STE4/100 PV panel using PSOGWO and GWO strategies for SDM is illustrated in Fig. 11. The best solution values are 3.0574E-4 and 6.0221E-4, respectively, for PSOGWO and GWO strategies.

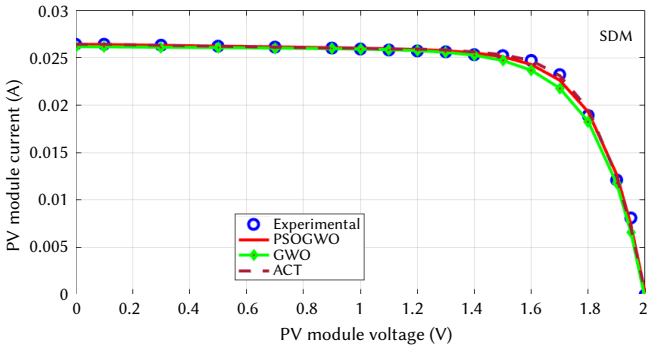



Fig. 9. The experimental dataset versus the estimated for STE4/100 PV panel.

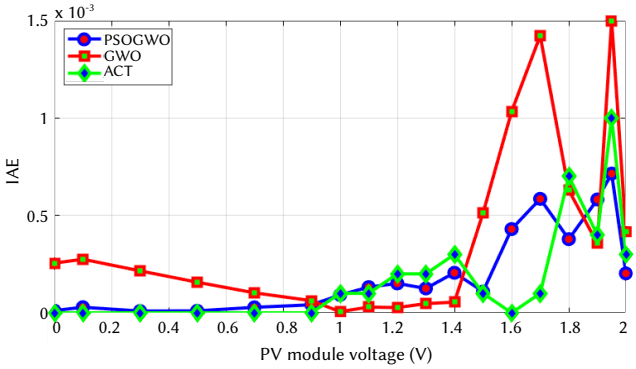



Fig. 10. Absolute error against measured PV module voltage of SDM for STP4/100 PV panel using different strategies.

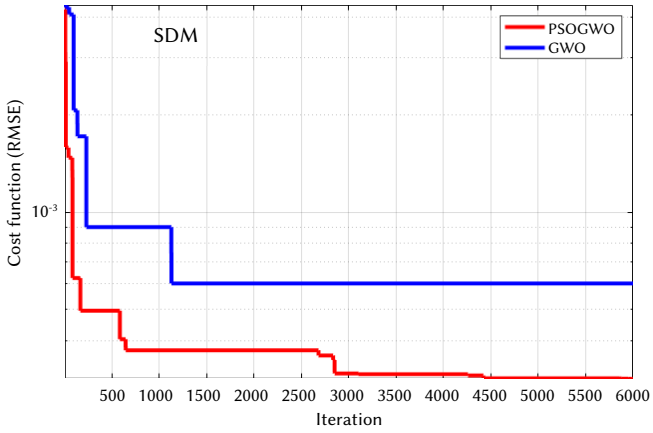



Fig. 11. The variation cost function during parameter estimation of STP4/100 PV panel using PSOGWO and GWO strategies.

The results of the whiteness test for STE4/100 PV panel using both PSOGWO and GWO strategies are shown in Fig. 12. It is very clear that the RCAF values range from -1 to +1 for both strategies.

### C. Results of 3rd Dataset

Based on the experimental dataset on FSM solar module, the proposed strategy of PSOGWO is used to determine the optimal parameters of the cell for both SDM and DDM. Table IV shows the maximum and minimum boundaries of each unknown parameter and optimal values of SDM and DDM parameters of FSM PV solar module using PSOGWO and GWO strategies.

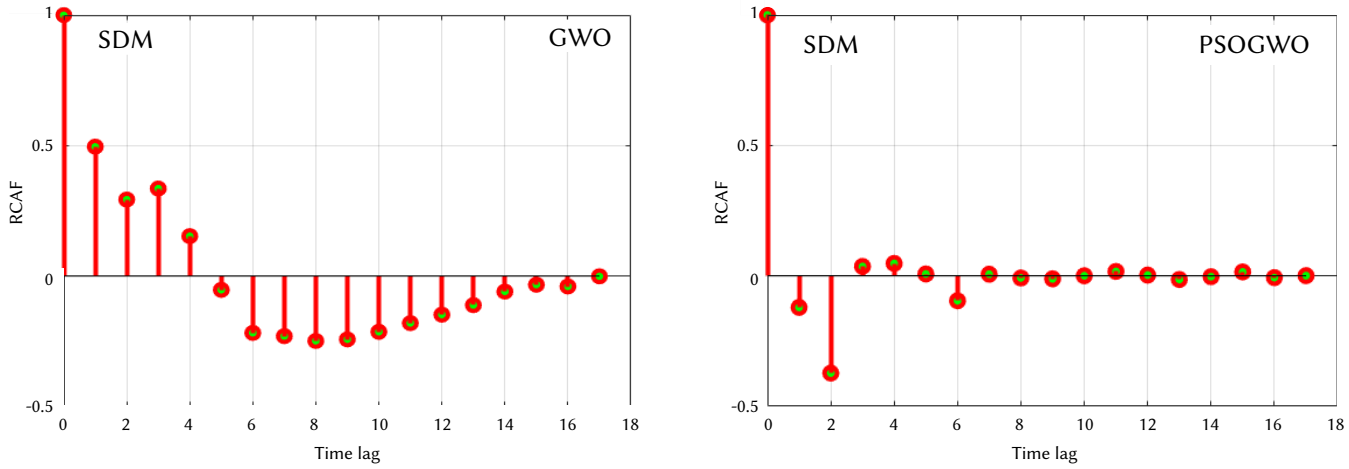



Fig. 12. RCAF results for STE4/100 PV panel using PSOGWO and GWO strategies.

TABLE IV. Boundaries and Optimal Values of SDM and DDM Parameters of FSM PV Module Using PSOGWO and GWO Strategies

| Parameter | Boundary | | Optimal parameters | | | |
|---|---|---|---|---|---|---|
| | Min. | Max. | SDM | | DDM | |
| | | | PSOGWO | GWO | PSOGWO | GWO |
| $I_{sc}$ (A) | 0 | 2 | 1.1132 | 1.1315 | 1.11027 | 1.12253 |
| $Io_1$ (A) | 1.0E-8 | 1.0E-3 | 1.012E-04 | 1.819E-04 | 5.40E-09 | 1.07E-06 |
| $Io_2$ (A) | 1.0E-8 | 1.0E-3 | na | na | 1.02E-04 | 1.081E-03 |
| $a_1$ | 0.0 | 3.0 | 2.232 | 2.275 | 1.8583 | 2.9803 |
| $a_2$ | 0.0 | 3.0 | na | na | 2.23125 | 2.9917 |
| $R_s$ (Ω) | 0.0 | 5.0 | 1.3357 | 2.211 | 1.36741 | 0.9787 |
| $R_{sh}$ (Ω) | 0.0 | 5000 | 1430.439 | 4980.36 | 3918.684 | 4907.89 |
| RMSE | | | 9.14E-03 | 2.99E-02 | 8.97E-03 | 2.52E-02 |
| MAE | | | 7.63E-03 | 2.12E-02 | 7.38E-03 | 1.95E-02 |
| $R^2$ | | | 0.9991 | 0.9901 | 0.9991 | 0.9929 |

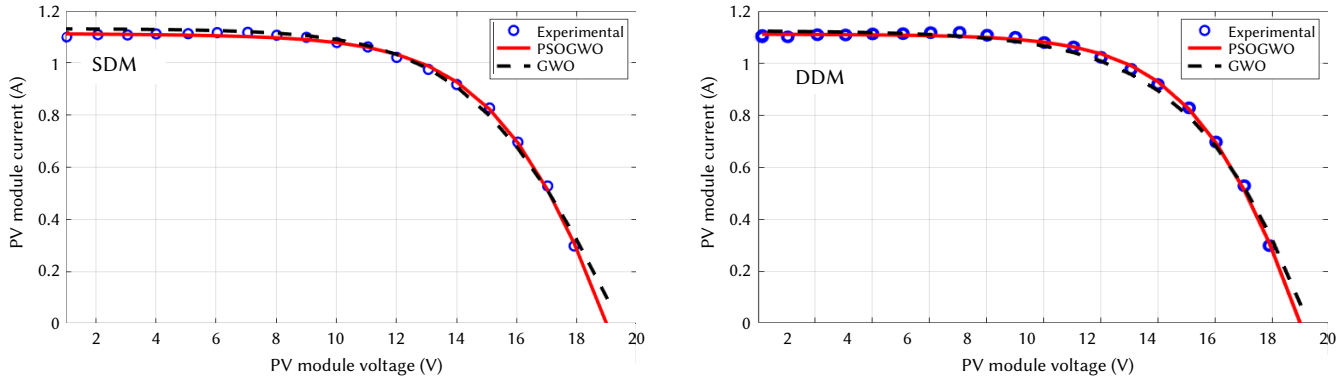



Fig. 13. The experimental dataset versus the estimated for FSM PV module for both SDM and DDM using PSOGWO and GWO strategies.

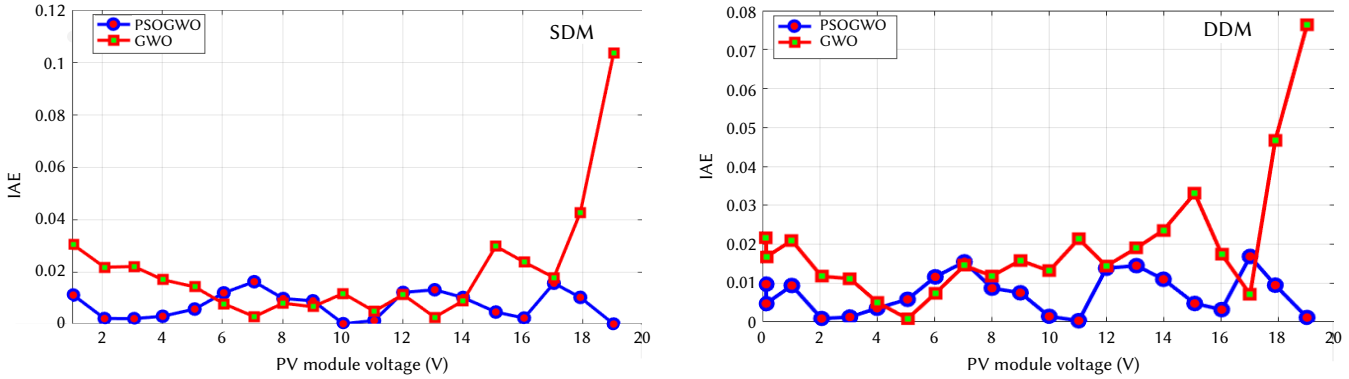



Fig. 14. Absolute error against measured PV module voltage for both SDM and DDM for FSM PV module using PSOGWO and GWO strategies.

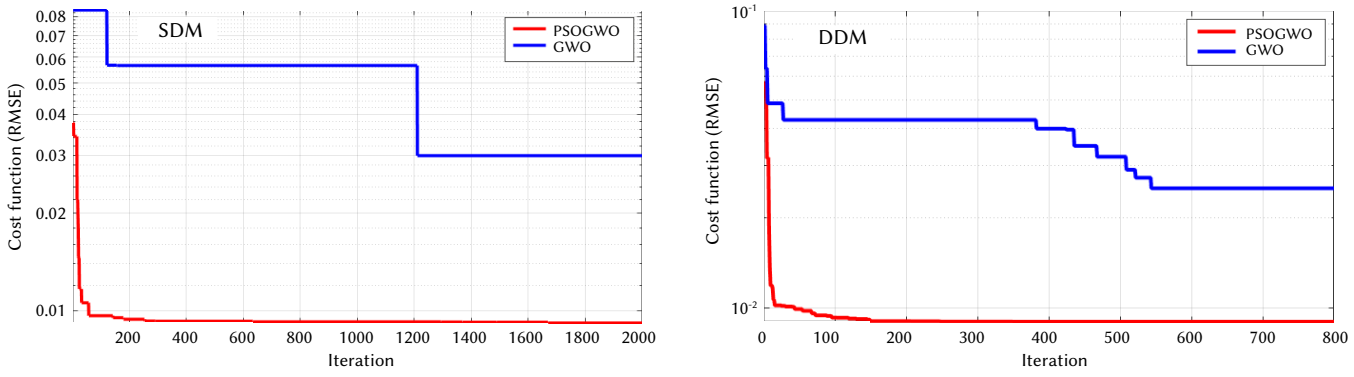



Fig. 15. The variation cost function during parameter estimation of FSM PV module using SFS strategy.

The minimum value of RMSE is achieved by PSOGWO strategy. For SDM, the values of RMSE are 7.63E-03 and 2.12E-02, respectively, for PSOGWO and GWO. Whereas, for DDM, the values of RMSE are 7.38E-03 and 1.95E-02 respectively for PSOGWO and GWO. The best coefficient of determination is 0.9991. It is achieved by using PSOGWO for both SDM and DDM. This confirms that there is an almost perfect agreement between the estimated datasets and the experimental data. Fig. 13 shows the experimental dataset versus the estimated, respectively, for both SDM and DDM.

The absolute error against measured PV module voltage for both SDM and DDM using PSOGWO and GWO strategies is shown in Fig. 14. For the SDM, the maximum values for the absolute error are 0.0162 and 0.1035, respectively, for PSOGWO and GWO. Whereas for the DDM, the maximum values for the absolute error are 0.0168 and 0.0764, respectively, for PSOGWO and GWO. This confirms the superiority of PSOGWO compared with GWO.

The variation of cost function during parameter estimation of FSM PV panel using PSOGWO and GWO strategies for both SDM and DDM is illustrated in Fig. 15. For the SDM, the best solution values are 9.14E-03 and 2.99E-02 respectively for PSOGWO and GWO strategies. Whereas for DDM as shown in Fig. 15(b), the best solution values are 8.97E-03 and 2.52E-02, respectively, for PSOGWO and GWO strategies. This confirms the superiority of PSOGWO compared with GWO for both SDM and DDM.

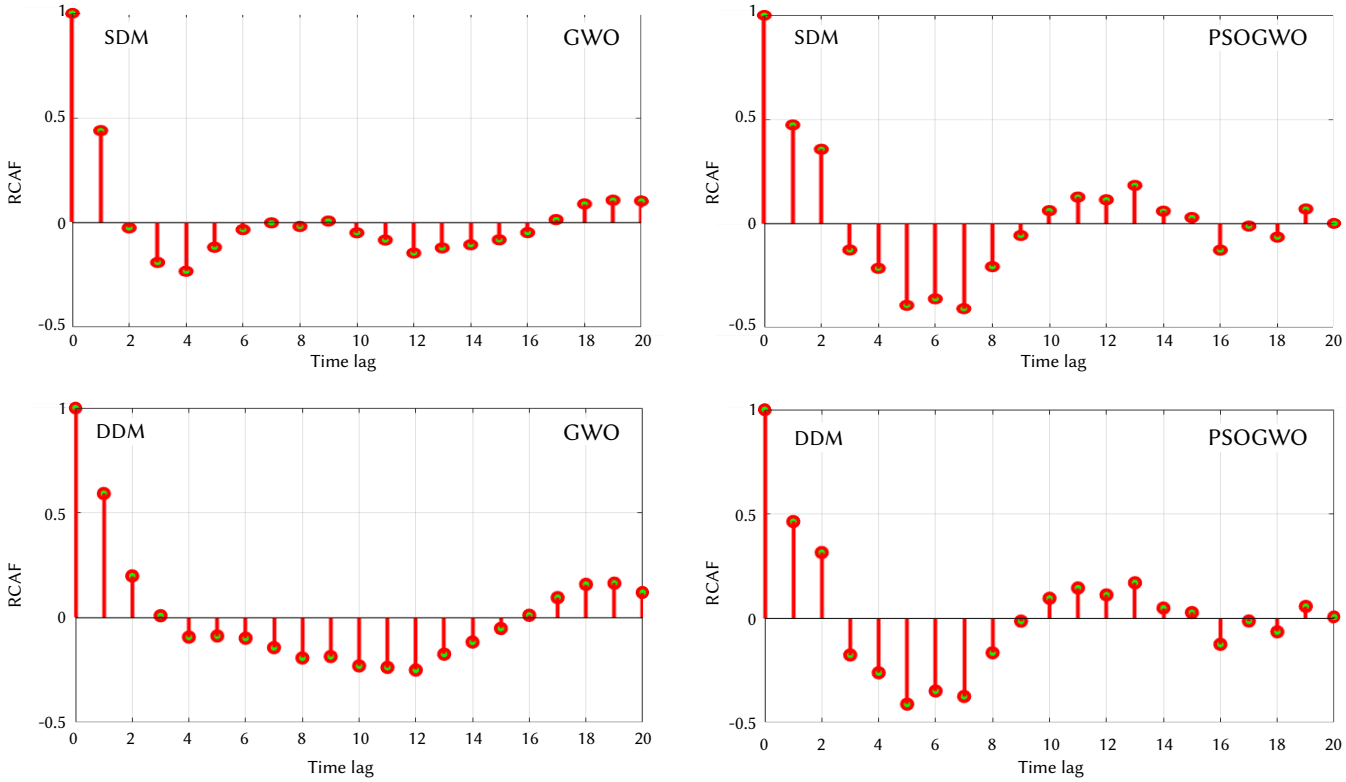



Fig. 16. RCAF results for FSM PV module using PSOGWO and GWO strategies.

The results of the whiteness test for STE4/100 PV panel using both PSOGWO and GWO strategies are shown in Fig. 16. It is very clear that the RCAF values range from -1 to +1 for both strategies.

## VI. Conclusion

Application of a hybrid particle swarm optimization (PSO) and grey wolf optimizer (WGO) in determining the optimal internal parameters of single-diode and double-diode models of a solar photovoltaic panel is presented for the first time in this paper, as far as the author know. Based on the experimental datasets of voltage-current curves, these internal parameters are determined. Three different PV panels are used to validate the propped strategy. The root mean square error, mean absolute error, and coefficient of determination are used as benchmark criteria for the comparison with other methods. For example, with the first dataset, in the case of SDM, the minimum value of RMSE is achieved by PSOGWO strategy. The values of RMSE are 3.06E-03, 1.43E-02, 1.18E-02, and 6.84E-03 respectively for PSOGWO, ALO-LW, PS, and GA. The coefficient of determination is 0.999952 using PSOGWO. This confirms that there is an almost perfect agreement between the estimated datasets and the experimental data. For all considered cases, the obtained results confirmed the superiority of PSOGWO compared with other methods.

## Acknowledgment

This project was supported by the Deanship of Scientific Research at Prince Sattam Bin Abdulaziz University under the research project No 2020/01/11742.

## References

[1] K. Ishaque, Z. Salam, H.Taheri. "Simple, fast and accurate two-diode model for photovoltaic modules," *Solar Energy Materials Solar Cells*, vol. 95, pp. 586-594, 2011.

[2] HGG. Nunes, JAN. Pombo, SJPS. Mariano, MRA. Calado and JAM. Felippe de Souza. "A new high performance method for determining the parameters of PV cells and modules based on guaranteed convergence particle swarm optimization," *Applied Energy*, vol. 211, pp. 774-791, 2018.

[3] J. Ma, Z. Bi, TO. Ting, S. Hao, W. Hao. "Comparative performance on photovoltaic model parameter identification via bio-inspired algorithms," *Solar Energy*, vol. 13, pp. 606-616, 2016.

[4] A. Ayang, W. René, O.Mohand, D. Noël, S. Ndjakomo Essiane, P. Joseph Kessel, and E. Gabriel . "Maximum likelihood parameters estimation of single-diode model of photovoltaic generator," *Renewable Energy*, vol. 130, pp. 111-121, 2019.

[5] A. Chatterjee, K. Ali, and K. Dhruv. "Identification of photovoltaic source models," *IEEE Transactions on Energy Conversion*, vol. 26, no.3, pp. 883-889, 2011.

[6] AK. Tossa, YM. Soro, Y. Azoumah, D. Yamegueu. "A new approach to estimate the performance and energy productivity of photovoltaic modules in real operating conditions," *Solar Energy*, vol. 10, pp. 543-560, 2014.

[7] T. Easwarakhanthan, J. Bottin, I. Bouhouch, C. Boutrit ."Nonlinear minimization algorithm for determining the solar cell parameters with microcomputers," *International Journal of Solar Energy*, vol.4, no.1, pp. 1-12, 1986.

[8] M. Chegaar, Z. Ouennoughi, A. Hoffmann. "A new method for evaluating illuminated solar cell parameters, " *Solid-State Electron*, vol.45, pp. 293-296, 2001.

[9] S.A. Blaifi, M. Samir, T. Bilal and S. Abdelhakim. "An enhanced dynamic modeling of PV module using Levenberg-Marquardt algorithm," *Renewable Energy*, vol. 135, pp. 745-760, 2019.

[10] S. Chen, F. Saeid Gholami and L. Sebastian. "Photovoltaic cells parameters extraction using variables reduction and improved shark optimization technique," *International Journal of Hydrogen Energy*, 2020.

[11] JA. Jervase, H. Bourdoucen, A. Al-Lawati. "Solar cell parameter extraction using genetic algorithms," *Measurement Science and Technology*, vol. 12, no.11, pp. 1922-1925, 2001.

[12] D. Oliva, A. A. Ewees, M. A. E. Aziz, A. E. Hassanien, M. Peréz-Cisneros. "A chaotic improved artificial bee colony for parameter estimation of photovoltaic cells," *Energies*, vol.10, pp. 865, 2017.

[13] R. Wang, Y. Zhan, H. Zhou. "Application of artificial bee colony in model parameter identification of solar cells," *Energies*, vol.8, no.8, pp. 7563-7581, 2015.

[14] M. Jamadi, F. Merrikh-Bayat, M. Bigdeli. "Very accurate parameter estimation of single- and double-diode solar cell models using a modified artificial bee colony algorithm," *International Journal Energy Environmental Engineering*, vol. 7, no.1, pp. 13-25, 2016.

[15] V. Khanna, B. K. Das, D. Bisht, Vandana, P. K. Singh. "A three diode model for industrial solar cells and estimation of solar cell parameters using PSO algorithm," *Renewable Energy*, vol. 78, pp. 105–113, 2015.

[16] R. Muralidharan. "Parameter extraction of solar photovoltaic cells and modules using current–voltage characteristics," *International Journal Ambient Energy*, vol.38, no.5, pp. 509-513, 2017.

[17] J. Ma, K. L. Man, S-U. Guan, T. O. Ting, P. W. H. Wong. "Parameter estimation of photovoltaic model via parallel particle swarm optimization algorithm," *International Journal of Energy Research*, vol.40, no.3, pp. 343-352, 2016.

[18] L. Guo, Z. Meng, Y. Sun, L. Wang ."Parameter identification and sensitivity analysis of solar cell models with cat swarm optimization algorithm," *Energy Conversion and Management*, vol.108, pp. 520-528, 2016.

[19] A. Askarzadeh, A. Rezazadeh. "Parameter identification for solar cell models using harmony search-based algorithms," *Solar Energy*, vol. 86, no.11, pp. 3241-3249, 2012.

[20] F. Dkhichi, B. Oukarfi, A. Fakkar, N. Belbounaguia. "Parameter identification of solar cell model using Levenberg–Marquardt algorithm combined with simulated annealing," *Solar Energy*, vol. 110, pp. 781-788, 2014.

[21] Z. Chen, L. Wu, P. Lin, Y. Wu, S. Cheng. "Parameters identification of photovoltaic models using hybrid adaptive Nelder-Mead simplex algorithm based on eagle strategy," *Applied Energy*, vol.182, pp. 47–57, 2016.

[22] N. F. Abdul Hamid, N. A. Rahim, J. Selvaraj. "Solar cell parameters identification using hybrid Nelder-Mead and modified particle swarm optimization," *Journal of Renewable and Sustainable Energy*, vol.8, pp. 1-21, 2016.

[23] X. Chen, B. Xu, C. Mei, Y. Ding, K. Li. "Teaching–learning–based artificial bee colony for solar photovoltaic parameter estimation," *Applied Energy*, vol. 212, pp. 1578-1588, 2018.

[24] K. Yu, B. Qu, C. Yue, S. Ge, X. Chen, J. Liang ."A performance-guided jaya algorithm for parameters identification of photovoltaic cell and module," *Applied Energy*, vol. 237, pp. 241-257, 2019.

[25] D. Oliva, M. A. El Aziz, A. E. Hassanien. "Parameter estimation of photovoltaic cells using an improved chaotic whale optimization algorithm," *Applied Energy*, vol. 200, pp. 141-154, 2017.

[26] A. M. Beigi, A. Maroosi. "Parameter identification for solar cells and module using a Hybrid Firefly and Pattern Search Algorithms," *Solar Energy*, vol.171, 2018.

[27] C. Saravanan, M.A. Panneerselvam. "A comprehensive analysis for extracting single diode PV model parameters by hybrid GA-PSO algorithm," *International Journal of Computer Applications*, vol.78, no.8, pp. 16-19, 2013.

[28] H. Nunes, J. Pombo, S. Mariano, M. Calado, J. F. de Souza. "A new high performance method for determining the parameters of pv cells and modules based on guaranteed convergence particle swarm optimization,"*Applied Energy*, vol. 211, pp. 774–791, 2018.

[29] D. Yousri, T. S. Babu, D. Allam, V. K. Ramachandaramurthy, M. B. Eteiba. "Fractional chaotic ensemble particle swarm optimizer for identifying the single, double, and three diode photovoltaic models' parameters," *Energy*, 2020.

[30] J. Kennedy and R. C. Eberhart. "Particle swarm optimization," *in Proceedings of IEEE international conference on neural networks*, 1995, vol.4, pp. 1942–1948.

[31] S. M. Mirjalili, A. Lewis. "Grey wolf optimizer," *Advances in Engineering Software*, vol. 69, pp. 46-61, 2014.

[32] G. K. Harish Kumar. "Modeling of solar cell under different conditions by ant lion optimizer with LambertW function," *Applied Soft Computing*, vol.71, pp. 141-151, 2018, doi : 10.1016/J.ASOC.2018.06.025.

[33] F. Fahmi, Muhammad and al. "Simple and efficient estimation of photovoltaic cells and modules parameters using approximation and correction technique," PLOS, 2019, doi: 10.1371/journal.pone.0216201.

[34] H. Rezk, I. Tyukhov, M. Al-Dhaifallah, A. Tikhonov ."Performance of data acquisition system for monitoring PV system parameters," *Measurement*, vol. 104, pp. 204-211.

[35] M. F. AlHajri, K. M. El-Naggar, M. R. AlRashidi, A. K. Al-Othman. "Optimal extraction of solar cell parameters using pattern search," *Renewable Energy*, vol. 44, pp. 238-245, 2012.

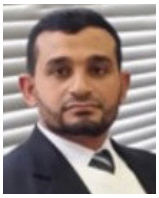
[36] M. R. AlRashidi, M. F. AlHajri, K. M. El-Naggar, A. K. Al-Othman. "A new estimation approach for determining the I–V characteristics of solar cells," *Solar Energy*, vol. 85, pp.1543-1550, 2011.

Hegazy Rezk

Hegazy Rezk received the B. Eng. and M. Eng. degrees in electrical engineering from Minia University, EGYPT in 2001 and 2006 respectively, and his PhD from Moscow Power Engineering Institute, Moscow. He was a postdoctoral research fellow in Moscow State University of Mechanical Engineering, Russia for 6 months. Dr. Hegazy was a visiting Researcher at Kyushu University, Japan, for one year. Currently, Hegazy Rezk is Associate Professor in Electrical Engineering Department, Collage of Engineering at Wadi Addwaser, Prince Sattam University, Saudi Arabia. He authored more than 100 technical papers. His present research interests include renewable energy, smart grid, hybrid systems, power electronics, Optimization and artificial intelligence.

Jouda Arfaoui

Jouda Arfaoui was born in Tunisia, in 1986. She received the Master's degree in Electronics from the Faculty of Sciences of Tunis in 2012. On May 2018, she received her PhD. Her research interests include renewable energy resources, fuzzy control of several applications, power system control and optimization.



Mohamed R. Gomaa

Mohamed R. Gomaa currently employing as Assis. Prof. of Thermal and Renewable energy, Mechanical Engineering at Benha Faculty of Engineering, Benha University, Banha, Egypt. I have awarded my PhD from SEUA (Polytechnic), Yerevan, Armenia on December 2012 in the field of Renewable Energy Systems. I had completed my MSc in the field of Fluid Mechanics on March 2007 from El-Minia University, Egypt. Currently, my research interests in the field of Thermal and Renewable energy systems, absorption, desalination and air-conditioning systems.

# Machine Learning Classifier Approach with Gaussian Process, Ensemble boosted Trees, SVM, and Linear Regression for 5G Signal Coverage Mapping

Akansha Gupta*, Kamal Ghanshala, R. C. Joshi

Graphic Era Deemed to be University, Dehradun (India)

Received 22 August 2020 | Accepted 4 January 2021 | Published 30 March 2021

## Abstract

This article offers a thorough analysis of the machine learning classifiers approaches for the collected Received Signal Strength Indicator (RSSI) samples which can be applied in predicting propagation loss, used for network planning to achieve maximum coverage. We estimated the RMSE of a machine learning classifier on multivariate RSSI data collected from the cluster of 6 Base Transceiver Stations (BTS) across a hilly terrain of Uttarakhand-India. Variable attributes comprise topology, environment, and forest canopy. Four machine learning classifiers have been investigated to identify the classifier with the least RMSE: Gaussian Process, Ensemble Boosted Tree, SVM, and Linear Regression. Gaussian Process showed the lowest RMSE, R- Squared, MSE, and MAE of 1.96, 0.98, 3.8774, and 1.3202 respectively as compared to other classifiers.

## Keywords

Propagation Loss, RSSI, Radio Propagation, Machine Learning Classifiers, 5G, SVM.

DOI: 10.9781/ijimai.2021.03.004

## I. Introduction

Wireless Networks have developed rapidly during the last 2-3 decades.1G arrived in the late 1980s which worked on analog signals and supported voice calls only. 2G arrived during the 1990s and was used for voice calls and data transmission having a bandwidth of 64kps.In 2000, 3G was launched with a bandwidth of 1Mbps to 2Mbps and supported not only voice calls but also video calls and conferencing.4G came into existence in 2009 with its data transmission speed of 100Mbps - 1Gbps.4G network was expanded world-wide and found its industrial applications also. Advance wireless technology is evolving very rapidly with the implementation of the 5G-NR network until 2020 [1]-[3]. Features of a High-speed next-generation 5G-NR network are high density (1 million nodes per Km2), high capacity (10Tbps per Km2), high data rates (Multi Gbps peak rates), low latency (1 ms), high reliability (1 out of 100 million packets lost), low energy (10 + years of battery life), low complexity, and high speed of 1 Gbps. With the establishment of this new technology, it may bring some challenges and difficulties like security, privacy, etc. Millimeter (MM) waves required for 5G propagation have a limitation of their effects on human cells and tissues, getting absorbed during transmission, require a small size antenna and cause unpredictable loss of signal during propagation [4]-[6].

In advanced wireless networks remarkable enhancement in information is observed [7]-[9]. Manual extraction of relevant information from an enormous amount of data is not possible, and if done, it will be prone to inevitable flaws. For capturing such big data companies no longer limit themselves to surveys and questionnaires rather big data capturing devices are deployed which include smart phone's, cameras, online browsing, etc. Machine learning seems to hold a promising solution for the analysis of big data [10]-[13]. Generally, data patterns are learned using information hidden in the big data, and then effective predictions can be proposed depending upon the final analysis. There are many Machine Learning algorithms available out of which appropriate selection of machine learning algorithms can be done using the hit and trial method [14]-[16]. We briefly describe the content in the following sections of the paper. In Section II and III a literature review and an outline of Machine learning algorithms is given respectively. In Section IV we describe measurement setup with data collection methodology. Finally, in Section V and VI, we report experimental results, discussions, conclusions, and future scope which clearly show the usefulness of the machine learning approach in predicting signal coverage.

* Corresponding author.

E-mail address: akanksha3000@gmail.com

## II. Literature Review

Hajar El Hammouti et al. [17] proposed a signal mapping model based on field measured data which is applied for predicting signal coverage in an outdoor network by utilizing an S-shaped sigmoid function. Effectively modeled neural networks provide a better approximation of coverage mapping. Amir Ghasemi [18] presented a crowd-sourced analysis of the Long-Term Evaluation (LTE) network to build a wireless coverage predictive model of the radio access network (RAN). Janne Riihijarvi et al. [19] explored signal coverage mapping with machine learning algorithms using a data set derived from an extensive drive test and analyzed that Random Forest, Exponential smoothing of time series, and Gaussian Process are machine learning methods that produce better results for signal coverage. It improves the Quality of user experience and concurrently reduces

the operational cost. H. Braham et al. [20] analyzed that accurate evaluation of coverage gives better coverage optimization by utilizing the Fixed Rank Kringing (FRK) algorithm which further provides an accurate prediction of signal coverage of the locations where field measurements are not easily accessible. FRK frames a coverage map from geo located measurement by interpolating them spatially. Carlos Oroza et al. [21] estimated the performance of a machine learning based signal loss model for different terrain and vegetation environments. Four major machine learning algorithms were explored with minimum error value: K-Nearest-Neighbor, Adaboost, Random Forest, and Neural Networks. Random Forest outperforms among them with the least error. Machine learning model accomplishes a 37% reduction in average prediction error. Many researchers [22]-[23] executed exhaustive field signal measurement on Received Signal Strength Indicator (RSSI) transmission multiband channel to get the maximum signal coverage and angular power arrival. A hybrid approach adopting sub 6-G and Millimeter (mm) wave bands was found to be promising. In [24]-[25] authors proposed Backtracking Spiral Algorithm (BSA) information detection and recognition of cellular coverage using the big-data method. The distribution potential of the individual cell was diagnosed in a small granular geographical grid. The high efficiency and detection capability of the proposed algorithm validate it over other existing algorithms. Aldebaro Klautau et al. [26] represented 5G scenarios by creating channel realizations using ML applied to the PHY layer. Tadilo Endeshaw et al. [27] studied the deployment of AI by combining machine learning, NLP, and data analytics techniques for increasing the competence of wireless networks. Deussom Djomadji et al. [28] tuned the propagation models using Particle Swarm Optimization. Data is collected using the network navigation tool IX EVDO rev B. Comparison of RMSE has been done for the optimized model and the Okumura Hata model and it was concluded that an optimized model using PSO performs better than Okumura Hata model. Chao-Kai Wen et al. [29] estimated the wireless channel based on Sparse Bayesian Learning techniques.

## III. Outdoor Path Loss Prediction

5G Network planning consists of outdoor propagation modeling required to predict the outdoor signal loss. Outdoor channel models consider the effect of diffraction, reflection, scattering, and refraction of EM waves, when traveling in free space between transmitter and receiver [38], [47]-[49]. Different channel models have been designed to consider the interference effect due to terrain, the height of a building, vegetation, rain, terrain, etc. Ideal path loss is predicted using the free-space path loss equation. Path loss models have been classified into canonical, empirical, deterministic, and stochastic propagation models [30]-[32].

### A. Free Space Path Loss Model

It is based on the Friis transmission equation and one of its simplest kinds of first-order approximation connectivity models. EM waves travel without any interference in Line of Sight (LOS) in free space. Receive Signal Strength (RSS) decreases as the square of the distance between Tx-Rx increases with a single path after neglecting the ground effect [33]-[35].

The free space path loss model is expressed in equation (1):

$$\frac{P_r}{P_t} = \frac{G_t G_r \lambda^2}{(4\pi)^2 d^2 L} \quad (1)$$

where,

$P_t$ = transmitted power ($dB$)

$P_r$ = power received by receiver ($dB$)

$G_r$ = gain of receiving antenna

$G_t$ = gain of transmitting antenna

$\lambda$ = signal wave length ($m$)

d = distance between Tx-Rx

L = system loss factor (L=1 for FSL)

Channel loss (dB) can also be expressed using equation (2).

$$L(dB) = 32.44 + \log(d) + 20\log(f) \quad (2)$$

### B. Machine Learning Algorithms

Machine learning algorithms are classified as [36], [39]-[42]:

***Supervised Learning***: This algorithm learns from the specimen data and related results to predict the correct result when given new data similar to specimen data.

***Unsupervised Learning***: This algorithm learns from the specimen data without any related results where the algorithm itself has to determine underlying data patterns and groups or cluster data based on some kind of similarity in their features.

***Reinforcement Learning***: This algorithm learns from specimen data that lack labels where the result or outcome is rewarded or penalized. It is like learning by trial & error.

The below described algorithms are supervised learning algorithms.

#### 1. Linear Regression

Linear regression is the most suitable machine learning algorithm for prediction[18].In linear regression, a set of independent input parameters(x) are considered to determine the output parameter(y) and there exists an association between the input and output parameter which is expressed in the form of the linear equation as:

$$y = wx + \epsilon$$

Where $\epsilon$ = intercept on y-axis

$w$ = slope of the line

The linear regression algorithm tries to find the best fit line by minimizing the root mean square error (RMSE) between true and predicted value.

#### 2. Support Vector Machine (SVM)

Support Vector Machine (SVM) can perform both classification and regression of the data. The data is classified into one of the groups by finding a hyperplane that divides the input instances into two classes. The input vectors located on the hyperplane are the support vectors. In cases where the input data is not linearly separable, suitable kernel functions are applied which map data into higher dimensions where data can be easily classified.

#### 3. Gaussian Process

Gaussian regression process is a non-parametric Bayesian approach that computes the probability distribution over permissible functions in the data [29]. The posterior probability is deduced from the prior distribution and the data. For a linear function

$$y = wx + \epsilon$$

The posterior probability is obtained using Baye's rule, equation (3):

$$p(w/y, X) = \frac{p(y/X, w)p(w)}{p(y/X)}$$

$$\text{Posterior Probability} = \frac{\text{likelihood x prior probability}}{\text{marginal likelihood}} \quad (3)$$

To make predictions at some random point x* one needs to calculate the predictive probability where a weighted measure of all the posterior probability distribution is evaluated using equation (4).

$$p\left(f^*/_{x^*}, y, X\right) = \int p\left(f^*/_{x^*}, w\right) p\left(w/_{y,X}\right) dw \quad (4)$$

Predictive probability is computed from posterior probability so that uncertainty measurements on the predictions can be provided by calculating their mean and variance.

### 4. Ensemble Learning

Sometimes a single machine learning algorithm is not able to provide the desired results, the expected result can be obtained by combining available algorithms [43]. The final result can be calculated by voting or averaging the result of individual algorithms. Major ensemble algorithms used are Bagging and Boosting.

#### a) Ensemble Bagging Tree with Random Forests

In Bagging firstly initial data set is utilized to reproduce a replica of the training set by using the Bootstrap sampling method. The bootstrap sampling method is used to create random samples from the initial data set where the sample size used as a training set is the same as the initial data set [44]. The random sample is generated from initial data by duplicating some sets of data multiple times and some records are not even considered once. The test data set is the initial and random sample sets that are used as the training data set. Secondly, multiple models are built from the training sets when the same algorithm is applied to them. Random Forests is a very good example to represent bagging benefits. In Random Forests, the best feature is selected for classification that converges the algorithm faster to a unique result. When the same data set is used a similar tree structure with associated prediction is obtained. Whereas the random forest after every split, bagging provides a random set of features for classification that probably result in the negligible association among classifications from sub-models.

#### b) Ensemble Boosting Tree with AdaBoost

AdaBoost implies Adaptive Boosting. Bagging works on 'simple voting' where each model is developed independently to provide an outcome [45]. The final result is obtained after analyzing the majority outcomes of the parallel ensemble. Boosting works on 'weighted voting' where each model provides an outcome that is based on majority selection. The final result is obtained by generating a sequential ensemble where greater weights are designated to the instances of preceding models that are misclassified. In every iteration, a model is built by rectifying the misclassification of the preceding model until no further corrections are required [46].

## IV. Experimental Setup and Data Collection

This research has been carried out in Dehradun, Uttarakhand-India which lies in the Himalayan ranges. Its geographical coordinates are within latitude 78.0322° E, longitude 30.3165° N. Uttarakhand is often regarded as a terrain full of the tree canopy and suburban environment with a combination of mountain, forest, residential building, commercial complex(2 to 6 storied), and free space. Since 2018 exhaustive field measurement has been carried out at fringe areas of Uttarakhand to measure the effect of the tree, forest, mountain, snow, and buildings on propagation loss.4 clusters (6 BTS each) were identified, covering a 36 Km$^2$ area shown in Fig. 1. These BTS are strategically chosen to cover RSSI to the tree canopy and mountain canopy. At each test point, exhaustive measurements were carried out repeatedly to calculate changing average RSSI signal variation due to environmental conditions. Hours of driving and route tracking were carried out around identified BTS to collect RSSI samples using CATIA and TEMS navigation tools as shown in Fig. 3. When the dense forest area started, RSSI samples became inaccessible.

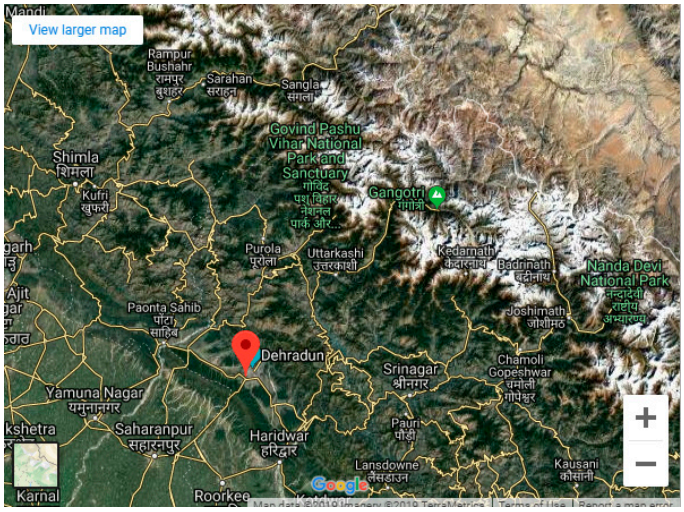



Fig. 1. The 4 clusters (6 BTS each) in the Dehradun,Uttarakhand-India.

A drive test was conducted to measure the data. As shown in Fig. 2 and Fig. 3 drive test tools consist of drive vehicle, laptop, Garmin (GPS) global position system, sockets, test cables, Sony mobile handset equipped with TEMS navigation software, MapInfo or Deskcat, and drive test route [35],[28].

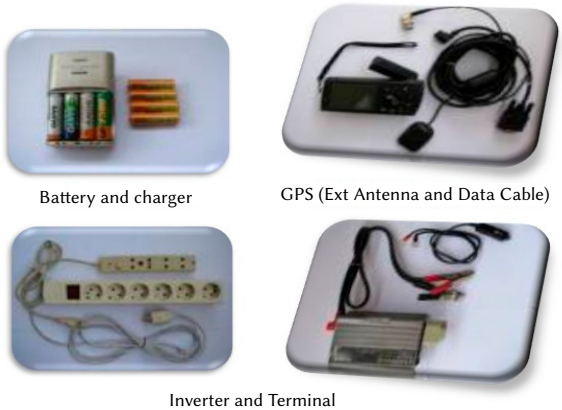



Fig. 2. 3D Drive test tool.

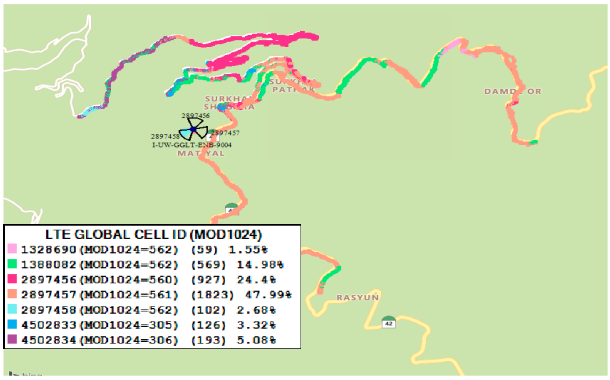



Fig. 3. Drive test route. Each BTS site is identified by a 3-letter codename.

GPS was mounted on a vehicle and a Sony Ericsson mobile handset was used.RSS measurements were recorded at each test point around the selected base stations (BS) covering all roads, forests, mountains, and populated areas.

### A. Dataset

42,500 RSSI samples of field measurement dataset are utilized for applying the machine learning approach on signal coverage prediction.

Fig. 4 shows the architecture of 3D channel modeling and the complete procedure of real-time data collection.

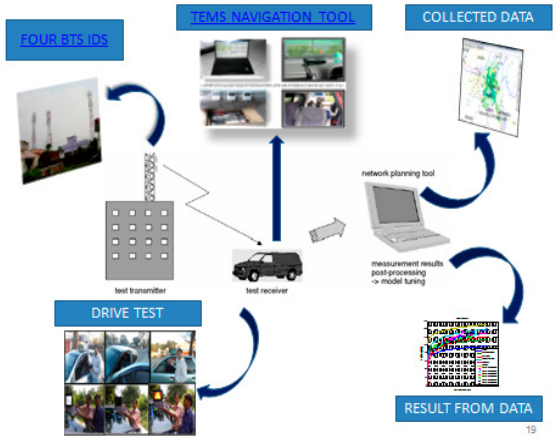



Fig. 4. Illustration of 3D Channel Modeling Architecture.

TABLE I. Features of Wireless Network

| Features | Value |
|---|---|
| **Coverage Objective RSRP (dBm)** | -106.6 |
| **Cluster Sectors** | I-UW-GGLT-ENB-9004-0 (A) |
| **Coverage Overshooting Radius (m)** | 4095 |
| **Band** | 850,2300,1800 |
| **Antenna Longitude** | 10 |
| **Antenna Latitude** | 50 |
| **Antenna Height (m)** | 37 |
| **Antenna Azimuth** | 10 |
| **Antenna Tilt Electrical** | 8 |
| **Antenna Tilt Mechanical** | 3 |

Table I illustrates features of wireless network which affect network planning and required for optimum signal coverage. The received signal is a combination of signals coming from different directions due to reflection, diffraction, and scattering.

RSSI signal strength is measured 360 degrees around each BTS in 3 sectors alpha, beta, and gamma to analyze the maximum coverage of signal within a cell. The coverage threshold values of the network in 3 sectors are summarizes in Table II.

TABLE II. Coverage Objective Threshold Value in Alpha, Beta, Gamma Sectors

| Coverage Objective Threshold | Sector Alpha, Beta, Gamma |
|---|---|
| **Coverage Objective RSRP (dBm)** | -106.6 |
| **Coverage Objective Percentile (%)** | 80 |
| **Coverage Overshooting Radius (m)** | 4095 |
| **Coverage Overshooting RSRP (dBm)** | -91.6 |
| **Coverage Overshooting Percentile (%)** | 10 |
| **Coverage Swap Percentile (%)** | 50 |
| **Coverage SideLobe Percentile (%)** | 30 |
| **Coverage Radius Inner Percentile (%)** | 10 |

## V. Experimental Results and Discussions

In this section, the performance of Machine learning classifiers is evaluated using data collected from the experimental setup. The validation scheme has been chosen before tuning to estimate the performance of the model on new data. Validation also helps to examine the predictive accuracy of the fitted models and avoids over fitting. 3 types of validation schemes were available:

a) Cross-Validation is used for small data sets and uses a full portion of the data set.

b) Hold out Validation is used for large data sets and uses some portion of the data set.

c) No Validation signifies no protection against overfitting.

5 fold cross-validation was used to divide the original data set into 5 disjoint sets as by using 5 fold cross-validation, the predictive accuracy of trained models was well estimated on the entire data set where each fold:

a) Trains a model

b) Evaluate the performance of model

c) Calculates average test error

### A. Predicted Vs Response Plot

The Predicted Vs Response plot analyzed the performance of classifiers by evaluating the efficiency of the regression model by investigating the prediction for varying response values. The predicted response of models was laid against the true response. An efficient regression model had a predicted response nearly identical to the true response, therefore response values lay close to the diagonal line. The perpendicular separation between the diagonal line to each point was the deviation of the prediction for the point under consideration. An efficient classifier has minimum errors and points distributed roughly identical about the diagonal line.

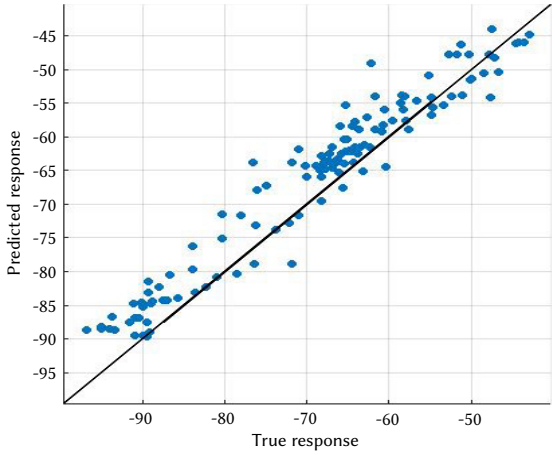



Fig. 5. Predicted Vs True response of Ensemble Boosted Tree.

Fig. 5 depicted RMSE value for Ensemble was 4.692 with R Squared value of 0.89. MSE was 22.015, MAE 3.88209, prediction speed 1800 obs/Sec, and training time of 32.441 Sec.

RMSE for Support Vector Machine was 3.9823 with R Squared value of 0.92, MSE 15.858, MAE 3.1148, prediction speed 11000 obs/Sec, and training time of 38.888Sec as shown in Fig. 6.

RMSE value for Linear Regression was 4.2946 with R Squared value of 0.91, MSE 18.444, MAE 3.5477, prediction speed 4000 obs/Sec, and training time of 25.623 Sec as shown in Fig. 7.

Fig. 8 shows RMSE value for Gaussian Process was 1.9691 with R Squared value of 0.98, MSE 3.8774, MAE 1.3202, prediction speed 8600 obs/Sec, and training time of 38.888 Sec.

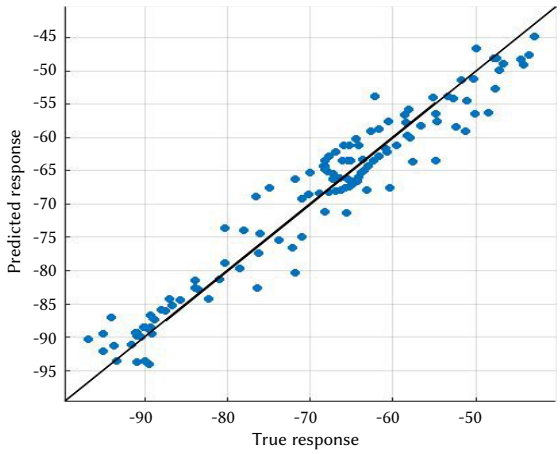



Fig. 6. Predicted Vs True response of Support Vector Machine.

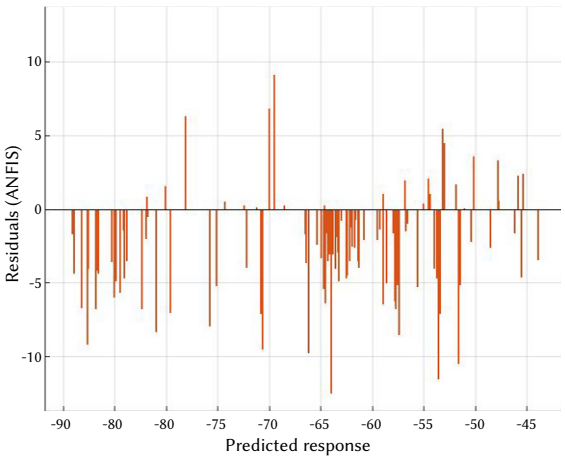



Fig. 9. Residual plot of Ensemble.

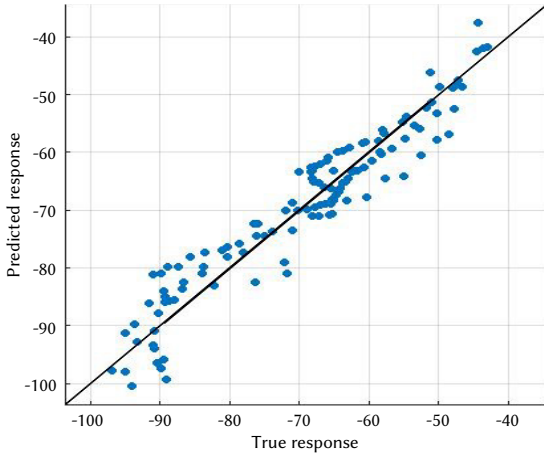



Fig. 7. Predicted Vs True response of Linear Regression.

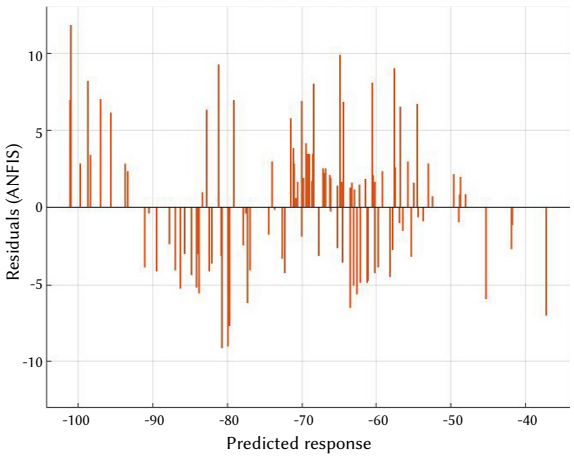



Fig. 10. Residual plot of Linear Regression.

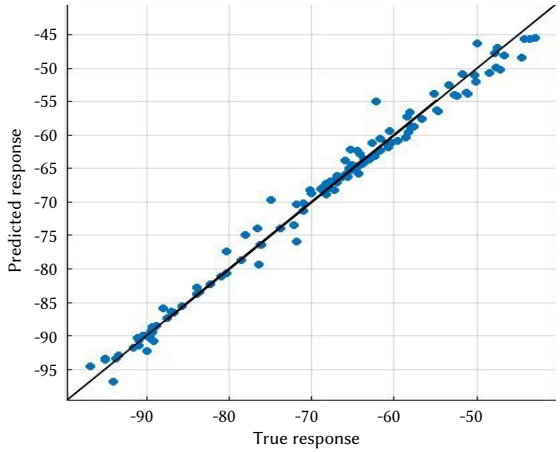



Fig. 8. Predicted Vs True response of Gaussian Process.

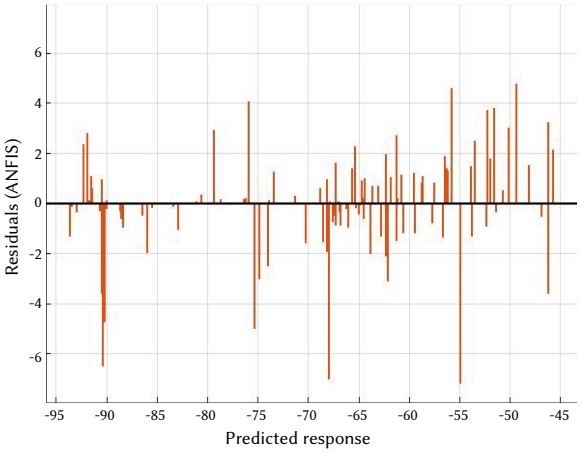



Fig. 11. Residual plot of Gaussian Process.

## B. Comparison of Residual Plot Outlier

The residuals plot from Fig.9 to Fig. 12 displayed the deviation between the predicted and true responses. The predicted response variable was chosen among true response, predicted response, record number, or one of the predictors to plot on the x-axis. The efficient model had residuals distributed roughly symmetrically around 0. Fig.10 showed that the residual plot of Gaussian Process scattered roughly symmetrically around 0 and also clear patterns in the residuals are observed.

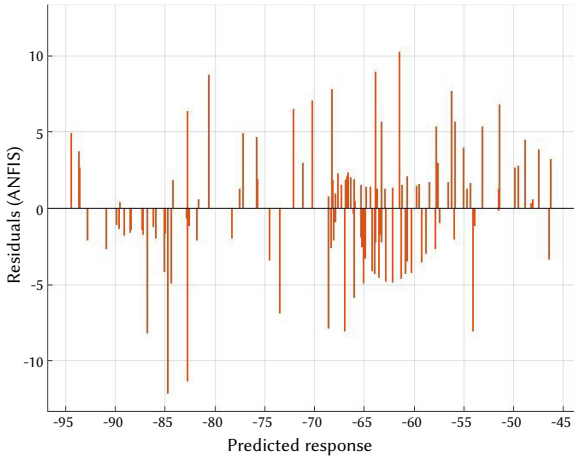



Fig. 12. Residual plot of SVM.

### C. Comparison of Response Plot Outlier

A regression model result was viewed in the response plot that displayed the prediction response against the record number. Predication error was displayed as vertical lines between predicted and new responses.

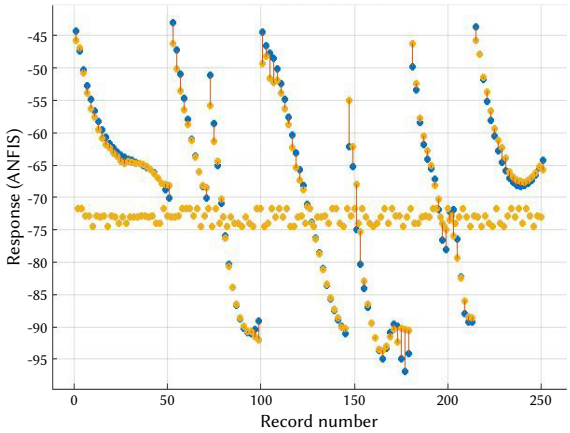



Fig. 13. Response plot of Gaussian Process.

Gaussian Process model performance was also evaluated using the residual plot after the model was trained. The difference between true and predicted response was displayed by a residual plot in Fig. 13. A variable predicted response had been plotted on the x-axis. For a good model, residuals had been scattered approximated identical around 0 and it changed considerably in size from left to right.

## VI. Conclusion and Observation

The performance evaluation of machine learning algorithms on training data is tabulated in Table III.

We observed that the lowest value of RMSE was obtained from the Gaussian Process classifier, depicting the probability to correctly predict the propagation loss, while the highest value of RMSE (4.692%) was observed with the Ensemble boosted tree. However, SVM and linear regression classifiers hold intermediate values of RMSE 3.9823%, and 4.2946%, respectively. The lowest value of RMSE (1.9691%) was estimated by the Gaussian Process classifier. MSE and MAE for Gaussian Process are also a minimum of 3.8774 and 1.3202. Response plot Outlier curves for the proposed Gaussian Process classifier and other state-of-the-art algorithms are shown in Fig. 13 where residuals have been scattered approximately identical around 0. It is analyzed that the Empirical signal coverage models which are univariate cannot predict signal coverage by using only one network parameter for coverage prediction, however machine learning-based signal coverage prediction model is multivariate and it could be designed on field RSSI measurement by considering two or more network parameters, hence predict signal coverage more accurately. Signal coverage prediction using the machine learning model requires training of best-fit machine learning classifier by hit and trial method and shortlisting machine learning classifiers with minimum RMSE error on RSSI field dataset. To validate it practically, the classifier-based signal mapping approach was applied to a real-time wireless network at the fringe area of Uttarakhand-India. However, the results of this application could encourage practitioners and researchers to validate further the practicality of the approach for similar real fringe area wireless networks.

TABLE III. Performance Evaluation of Machine Learning Algorithms on Training Data

| Coverage Objective Threshold | Gaussian Process | Ensemble Boosted Trees | SVM | Linear Regression |
|---|---|---|---|---|
| RMSE | 1.9691 | 4.692 | 3.9823 | 4.2946 |
| R Squared | 0.98 | .89 | 0.92 | 0.91 |
| MSE | 3.8774 | 22.015 | 15.858 | 18.444 |
| MAE | 1.3202 | 3.8209 | 3.1148 | 3.5477 |
| Prediction Speed (obs/Sec) | 8600 | 1800 | 11000 | 4000 |
| Training Time (Sec) | 38.888 | 32.441 | 26.495 | 25.623 |

## References


[1] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, V. Young, "Mobile edge computing-A key technology towards 5G," in ETSI White Paper, vol. 11, no. 11, pp. 1–16, Sept. 2015, ISBN No. 979-10-92620-08-5.

[2] K. Cichon, H. Bogucka, G. Molis, J. Adamonis, T. Krilavicius, "Learning and detection mechanisms of spectral-activity information towards energy efficient 5G communication," Baltic URSI Symposium (URSI), Poznan, Poland, May 2018,doi: 10. 23919/URSI. 2018.8406715.

[3] Y. Liu, Y. Zhang, R. Yu, S. Xie, "Integrated energy and spectrum harvesting for 5G wireless communications," in IEEE Network, vol. 29, no. 3 ,pp. 75 – 81, 2015,doi: 10. 1109/MNET. 2015.7113229.

[4] J. Kim, A. F. Molisch, "Fast millimeter-wave beam training with receive beamforming," in Journal of Communications and Networks, vol. 16, no. 5, pp. 512–522, Oct. 2014, doi: 10. 1109/JCN.2014.000090.

[5] E. Balevi, R. D. Gitlin, "Unsupervised Machine Learning in 5G Networks for Low Latency Communications," in proc. IEEE 36th international Performance computing and communication conference (IPCCC), ISSN: 2374-9628, pp. 1-2, 2017, doi: 10. 1109/PCCC.2017.8280492.

[6] T. Bogale, X. Wang, L. Le, "Machine Inteligence Techniques for Next-Generation Context-Aware Wireless Network," in ITU Journal: ICT Discoveries, vol. no. 1, pp. 1-11, Feb. 2018, doi:arXiv:1801.04223.

[7] W. Lee, "Mobile Communication Design Fundamentals", New York, John Wiely&Sons, 1993.

[8] B. Li, Z. Fei, Y. Zhang, "UAV Communications for 5G and Beyond: Recent Advances and Future Trends," in IEEE Internet of Things Journal, vol. 6, no. 2, pp. 2241 – 2263, April 2019, doi: 10. 1109/JIOT.2018.2887086.

[9] P. Antonia, A. Markos, T. Anna, S. Dimitra, "Provisioning of 5G services employing machine learning techniques," in International Conference on Optical Network Design and Modeling (ONDM) Dublin, Ireland, May 2018,doi: 10. 23919/ONDM. 2018.8396131.

[10] C. Jiang, H. Zhang, Y. Ren, Z. Han, K. C. Chen, L. Hanzo,"Machine learning paradigms for next-generation wireless networks," IEEE Wireless Communications., vol. 24, no. 2, pp. 98–105, 2017, doi: 10. 1109/MWC.2016.1500356WC.

[11] P. V. Klaine, M. A. Imran, O. Onireti, R. D. Souza, "A survey of machine

learning techniques applied to self-organizing cellular networks," in IEEE Communications Surveys and Tutorials, vol. 19, no. 4, pp. 2392–2431, 2017, doi: 10.1109/comst.2017.2727878.

[12] E. Bjorn, "Machine learning for beam based mobility optimization in NR," Master of Science dissertation, Department of Electrical Engineering, Linkoping University, Linkoping, Sweden, 2017.

[13] D. Krajzewicz, J. Erdmann, M. Behrisch, L. Bieker, "Recent development and applications of SUMO - Simulation of Urban Mobility," in International Journal On Advances in Systems and Measurements, vol. 5, no. 3&4, pp. 128–138, Dec. 2012, doi: 10.1.1.671.2113.

[14] J. G. Carbonell, R. S. Michalski, T. M. Mitchell, "An overview of machine learning," in book Machine learning, vol. 1, pp. 3-23, Elsevier Inc. 1983, doi: https://doi.org/10.1016/bs.host.2018.07.004

[15] M. Bkassiny, Y. Li, S. K. Jayaweera, "A survey on machine-learning techniques in cognitive radios," in IEEE Communications Surveys & Tutorials, vol. 15, no. 3, pp. 1136–1159, Oct. 2013. ,doi: 10.1109/SURV.2012.100412.00017.

[16] S. Osvaldo, "A Very brief introduction to machine learning with applications to communication systems," IEEE Transactions on Cognitive Communications and Networking, vol 4, no. 4, pp. 648 - 664 ,Dec. 2018, doi: 10.1109/TCCN.2018.2881442.

[17] H. Hajar, G. Mounir, Z. Syed, "A Machine Learning Approach to Predicting Coverage in Random Wireless Networks," in IEEE Globecom Workshops, pp. 1-6, 2018, doi: 10.1109/GCWkshps43968.2018.

[18] G. Amir, "Data-driven prediction of cellular networks coverage: an interpretable machine-learning model," in 2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 604-608, 2018, doi: 10.1109/GlobalSIP.2018.8646338.

[19] R. Janne,M. Petri, "Machine learning for performance prediction in mobile cellular networks," in IEEE Computational Intelligence Magazine, pp. 51-60, vol 13, Issue:1,2018,doi: 10. 1109/MCI.2017.2773824.

[20] H. Braham, S. B. Jemaa, G. Fort, E. Moulines, B. Sayrac, "Fixed rank kriging for cellular coverage analysis," IEEE Transaction Vehicular Technology, vol. 66, no. 5, pp. 4212–4222, May 2017. doi:10.1109/TVT.2016. 2599842.

[21] O. Carlos, Z. Ziran, W. Thomas, G. Steven, "A Machine-Learning based connectivity model for complex terrain large-scale low-power wireless deployments," IEEE Transactions on Cognitive Communications and Networking (Vol. 3, No. 4, Dec. 2017 ). doi: 10.1109/TCCN.2017.2741468.

[22] R. Nihesh, S. Renu, S. Rajesh, "Coverage estimation in outdoor heterogeneous propagation environments," in IEEE Access, vol. 8,pp. 31660 – 31673, doi: 10.1109/ACCESS.2020.2972811.

[23] Z. Zhimeng, Z. Jianyao, L. Chao, "Outdoor-to-Indoor channel measurement and coverage analysis for 5G typical spectrums," in International Journal of Antennas and Propagation, Vol. 2019, pp. 1-10, doi: 10.1155/2019/3981678.

[24] V. Kristem, S. Sangodoyin, C. U. Bas, "3D MIMO outdoor-to-indoor propagation channel measurement," IEEE Transactions on Wireless Communications, vol. 16, no. 7, pp. 4600–4613, 2017.

[25] W. Hai, X. Su, L. Ke, M. Omair, "Big data-driven cellular information detection and coverage identification" in sensor journal, vol 19, no. 4, pp. 937-942, 2019, doi: 10.3390/s19040937.

[26] K. Aldebaro, B. Pedro, G. Nuria, W. Yuyang, H. Robert, "5G MIMO data for machine learning: application to beam-selection sing deep learning," in Information Theory and Applications Workshop (ITA), San Diego, CA, USA, October 2018, doi: 10.1109/ITA.2018.8503086.

[27] B. Tadilo, B. Long, "Massive MIMO and mmWave for 5G wireless HetNet: potential benefits and challenges," IEEE Vehicular Technology Magazine, vol. 11, no. 1, pp. 64 – 75, March 2016, doi: 10.1109/MVT.2015.2496240.

[28] M. Deussom, E. Tonye, "New propagation model optimization approach based on Particles Swarm Optimization algorithm," in International Journal of Computer Applications, vol. 118, no. 10, pp. 39-47, 2015 ,doi: 10. 5120/20785-3430.

[29] W. Chao, J. Shi, W. Kai-Kit, C. Jung, T. Pangan, "Channel estimation for massive MIMO using gaussian-mixture bayesian learning," in IEEE Transaction Wireless Communication, vol. 14, no. 3, pp. 1356–68, Mar. 2015, doi: 10.1109/TWC.2014.2365813.

[30] S. Sun, "Path loss models for 5g urban micro- and macro-cellular scenarios," in 2016 IEEE VTC-Spring 2016, May 2016. [Online]. Available: http://arxiv. org/abs/1511.07311.

[31] M. Hall, "Radiowave propagation effects on next-generation fixed-services terrestrial telecommunication systems," European Cooperation in Science and Technology (COST), Tech. Rep. ICT COST Action 235, 1996.

[32] A. Gupta, S. Sharma, S. Vijay, V. Gupta, "Secure path loss prediction using fuzzy logic approach," in 2008 Fourth International Conference on Wireless Communication and Sensor Networks, Allahabad, India, Dec. 2008, doi: 10.1109/WCSN.2008.4772717.

[33] J. Azevedo, F. Santos, "An empirical propagation model for forest environments at tree trunk level," IEEE Transactions on Antennas and Propagation, vol. 59, no. 6, pp. 2357–2367, 2011.

[34] I. T. U. (ITU), "Influence of terrain irregularities and vegetation on tropospheric propagation," in CCIR XVth Plenary Assembly, vol. V: Propagation in Non-Ionised Media, no. ITUR Report 236-6, Dubrovnik, Croatia, 1986.

[35] S. Kim, B. Guarino, T. Willis, V. Erceg, S. Fortune, R. Valenzuela, L. Thomas, J. Ling, J. Moore, "Radio propagation measurements and prediction using three-dimensional ray tracing in urban environments at 908 MHz and 1. 9 GHz," in IEEE Transactions on Vehicular Technology, vol. 48, no. 3, pp. 931–946, May 1999, doi: 10.1109/25.765022.

[36] E. Alpaydm, "Introduction to Machine Learning," 3rd edition, The MIT Press, Cambridge, Massachusetts, 2014.

[37] S. Maghsudi, S. Stanczak, "Channel selection for network assisted D2D communication via no-regret bandit learning with calibrated forecasting," in IEEE Transactions on Wireless Communications, vol. 14, no. 3, pp. 1309–22, 2015, doi: 10.1109/TWC.2014.2365803.

[38] V. Gupta, S. Sharma, M. Bansal, "Fringe area path loss correction factor for wireless communication," in International Journal of Recent Trends in Engineering ,vol. 1, no. 2, pp. 30-32, May 2009, ISSN(online): 2455–1457.

[39] T. Shea, J. Hoydis, "An Introduction to Deep Learning for the Physical Layer," in IEEE Transactions on Cognitive Communications and Networking, vol. 3, no. 4, pp. 563–575, 2017,doi: 10.1109/tccn.2017.2758370.

[40] Y. LeCun, Y. Bengio, G. Hinton, "Deep learning," in Nature, vol. 521, pp. 436–444, 2015,doi: 10.1038/nature14539.

[41] R. Atallah, C. Assi, M. Khabbaz, "Deep reinforcement learning- based scheduling for roadside communication networks," in Proceeding of International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), Paris, France, pp. 1–8, May 2017, doi: 10.23919/wiopt.2017.7959912.

[42] T. Wang, C. Wen, H. Wang, F. Gao, T. Jiang, S. Jin, "Deep learning for wireless physical layer: Opportunities and challenges," in Journal China Communications, vol. 14, no. 11, pp. 92-111, Nov. 2017, doi: 10.1109/CC.2017. 8233654.

[43] R. Schapire, "The boosting approach to machine learning: An overview," in book Nonlinear Estimation and Classification, pp. 149–171, Springer, New York, NY, 2003, doi: 10.1007/978-0-387-21579-2.

[44] M. Prasad, L. Iverson, A. Liaw, "Newer classification and regression tree techniques: bagging and random forests for ecological prediction," in Springer International journal Ecosystems, vol. 9, no. 2, pp. 181-199, 2006, doi: 10.1007/s10021-005-0054-1.

[45] T. Shea, T. Erpek, T. C. Clancy, "Deep learning based MIMO communications," CoRR, vol. abs/1707. 07980, 2017. [Online]. Available: http://arxiv. org/abs/1707.07980.

[46] W. Loh, "Classification and regression trees," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 1, no. 1, pp 14–23, 2011, doi: 10.1002/widm.8.

[47] S. Salous, J. Kim, M. Sasaki, W. Yamada, X. Raimundo, A. Cheema, "Radio propagation measurements and modeling for standardization of the site general path loss model in International Telecommunications Union(ITU) recommendations for 5G wireless networks," In Radio Science Journal, (2020). vol 55, no. 1, Jan. 2020, doi: 10.1029/2019RS006924.

[48] S. Pattanayak, "A genetically trained neural network for prediction of path loss in outdoor microcell," in International Journal of Advanced Research in Engineering and Technology (IJARET), vol. 11, no. 4, pp. 346-351, 2020, doi: https://ssrn.com/abstract=3599786.

[49] A. Akinbolati, M. Ajewole, "Investigation of path loss and modeling for digital terrestrial television over Nigeria," in International Journal, Heliyon, 2020 Jun, vol. 6, no. 6, doi: 10.1016/j.heliyon.2020.e04101.

Akansha Gupta

Akansha Gupta received her B.Tech. degree in Electronics and Instrumentation Engineering from the UP Technical University, India, in 2005, and her M.Tech. degree in Computer Science Engineering from Uttarakhand Technical University India, in 2009. She is currently pursuing her Ph.D. degree in Computer Science engineering in developing future AI channel models for next generation 5G mobile cellular networks. Her research interests include Machine learning, wireless communication, IoT, 5G network, random matrix theory, and information theory.

Prof. (Dr.) Kamal Ghanshala

Prof. Kamal Ghanshala is an engineer, entrepreneur and a philanthropist with Bachelor's and Master's in Computer Science and Engineering. He has received his doctoral degree in Computer Science from Kumaon University, Nainital-India.He has received recognition for his research in conferences held at, Croatia, Denmark, Johannesburg, Turkey, London, Paris Germany and Thailand. He received the Visionary Edupreneur of India award 2017 from former president of India and handled many research projects. He has also received excellence award in the field of higher education in the international summit organized in Newyork, USA. He has founded two universities as a President at Uttarakhand-India, Graphic Era Deemed to be University and Graphic Era Hill University. His research interests center around the optimization in wireless multimedia networks, stochastic optimization method, and graphical approach for information processing.

Prof. (Dr.) R.C. Joshi

Dr. R.C. Joshi former Prof. E. & C.E. Department at IIT Roorkee and Chancellor at Graphic Era Deemed to be University Dehradun, received his B.E degree from NIT Allahabad in1967, M.E.1st Div. with Honors and Ph.D Degree from Roorkee University, now IIT Roorkee, in 1970 & 1980 respectively. He worked as a Lecturer in J.K Institute, Allahabad University during 1967-68. He had been Head of Electronics & Computer Engineering from Jan 1991-1994 & Jan. 1997 to Dec. 1999. He was also the Head of Institute Computer Centre, IIT Roorkee from March 1994 to Dec. 2005.He was on short visiting Professor's Assignment in University of Cincinnati, USA. University of Minnesota, U.S.A & Macquarie University Sydney Australia also visited France under Indo-France collaboration program during June 78 to Nov. 79. Dr. Joshi has guided 27 Ph.D, 250 M.Tech, Dissertation, 75 B.E Projects. He had taught more than 25 subjects in Computer Engineering, Electronics Engineering & Information Technology.

# Satisfiability Logic Analysis Via Radial Basis Function Neural Network with Artificial Bee Colony Algorithm

Mohd Shareduwan Mohd Kasihmuddin[1], Mohd. Asyraf Mansor[2]*, Shehab Abdulhabib Alzaeemi[1], Saratha Sathasivam[1]

[1] School of Mathematical Sciences, Universiti Sains Malaysia, 11800 USM, Penang (Malaysia)
[2] School of Distance Education, Universiti Sains Malaysia, 11800 USM, Penang (Malaysia)

Received 6 October 2019 | Accepted 7 May 2020 | Published 25 June 2020



## Abstract

Radial Basis Function Neural Network (RBFNN) is a variant of artificial neural network (ANN) paradigm, utilized in a plethora of fields of studies such as engineering, technology and science. 2 Satisfiability (2SAT) programming has been coined as a prominent logical rule that defines the identity of RBFNN. In this research, a swarm-based searching algorithm namely, the Artificial Bee Colony (ABC) will be introduced to facilitate the training of RBFNN. Worth mentioning that ABC is a new population-based metaheuristics algorithm inspired by the intelligent comportment of the honey bee hives. The optimization pattern in ABC was found fruitful in RBFNN since ABC reduces the complexity of the RBFNN in optimizing important parameters. The effectiveness of ABC in RBFNN has been examined in terms of various performance evaluations. Therefore, the simulation has proved that the ABC complied efficiently in tandem with the Radial Basis Neural Network with 2SAT according to various evaluations such as the Root Mean Square Error (RMSE), Sum of Squares Error (SSE), Mean Absolute Percentage Error (MAPE), and CPU Time. Overall, the experimental results have demonstrated the capability of ABC in enhancing the learning phase of RBFNN-2SAT as compared to the Genetic Algorithm (GA), Differential Evolution (DE) algorithm and Particle Swarm Optimization (PSO) algorithm.

## Keywords

Artificial Bee Colony Algorithm, Radial Basis Function Neural Network, 2 Satisfiability, Logic.

DOI: 10.9781/ijimai.2020.06.002

## I. Introduction

Generally, Artificial Intelligence (AI) can be encompassed in some functional graphical and mathematical models that act as a symbolic system [1]. Greater impact can be achieved when symbolic operations have been integrated inside the Artificial Neural Network (ANN). ANN known as the conductive system has received careful attention due to its ability to evaluate the complex nonlinear dataset [1]. ANN has been successfully used in solving non-limited applications such as classification and optimization of approximation functions. However, the functionality of ANN can be measured by embedding the correct symbolic rule to govern the whole neural system. Logic programming has been a language of ANN for decades. Wan Abdullah [2] successfully explored the neural network that has been governed by logic programming. In this work, logical rule embodied ANN and the characteristic of the network will be examined by using Lyapunov energy analysis. The minimization of energy as a solution to the combinatorial representation motivates the integration of logical rules in a neural network [3]-[6]. The question remains on how one can choose the best ANN model in order to embed logic programming.

The reliable ANN model typically has the least prediction and classification error analysis. In that regard, Radial Basis Function Neural Network (RBFNN) fascinated the researchers from sciences and engineering field because of simpler networks structure, faster learning speed and better approximation capabilities. As stated by Hamadneh *et al.* [7] in their paper, RBFNN can be used to develop separate models for the shear stress and heat transfer rate due to simpler networks structure. RBFNN is a feedforward neural network that contains 3 neuron layers (input, hidden and output layers). The input layer (containing input neurons) receives information being transferred to the hidden layer for data synthesis and training. The synthesized data will be used in the output layer (containing output neuron). The foundation of having 3 layers is to minimize the classification and prediction error in RBFNN [8]. Hamadneh *et al.* [4] initially implemented logic programming in RBFNN. Their proposed network explored the capability of HornSAT as a logical rule in RBFNN. In this case, logical structure of RBFNN is solely dependent on 3 parameters: the center of all input neurons, its widths, and its Gaussian activation function. Despite the fact that RBFNN can be applied effectively, the number of neurons in a hidden layer for RBFNN will determine the complexity of the network [9]. If the number of neurons in the hidden layers is not enough, the learning in RBFNN fails to achieve optimal convergence. However, if the number of neurons in the hidden layers is very high, the network will experience overlearning [10]. Since the complexity of RBFNN increases as the number of clauses increase, an optimization algorithm becomes crucial.

Yang and Ma [11] have successfully applied the Sparse Neural Network (SNN) algorithm for optimizing the number of hidden neurons. The core mechanism of SNN is in reducing the error via trial

* Corresponding author.
E-mail address: asyrafman@usm.my

and error approach for determining the number of hidden neurons explicitly from the set of neurons. The limitation of SNN paradigm can be seen in extensive computational time during the number of hidden neuron computation process. Inspired by several works of [12]-[14], the 2 Satisfiability (2SAT) logic representation will be utilized with RBFNN to determine the important parameters for the hidden layer that control the number of hidden neurons. In fact, 2SAT is selected since it is complying with RBFNN based on the structure and representations.

Another major component of 2SAT in RBFNN is the training method that has a significant influence on the performance of RBFNN. On this matter, a plethora of global optimization methods have been extensively applied due to their global search capability. Metaheuristics algorithm is a popular algorithm to search for a near optimal solution for RBFNN [15], [16]. There are various nature-inspired and recently developed optimization algorithms such as Genetic Algorithm, Differential Evolution algorithm, Particle Swarm Optimization algorithm, Artificial Bee Colony, etc. and many of these proved their suitability to many engineering optimization problems [17].

The theoretical basis of the Genetic Algorithm (GA) has been developed by Holland [18]. The first who used GA in a problem involving the control of gas-pipeline transmission were Goldberg and Holland [19]. Other studies have been made by Hamadneh *et al.* [4] who used GA to train the hybrid model RBFNN with higher-order SAT logic. In this study, they used the full training paradigm to train RBFNN with higher-order SAT logic using k-means cluster algorithm and GA. The quest of finding the optimal algorithm was continued by Pandey *et al.* [20] who compared Multiple Linear Regression (MLR) and genetic algorithm to predict temporal scour depth near-circular pier in non-cohesive sediment. This study utilized 1100 laboratory experimental data-sets to develop the generalized scour equation using MLR and GA. In recent publications, Jing and Li [21] developed a reliability analysis method by integrating GA with RBFNN. This paper adopted GA to find the "potential" most probable point (MPP) in the optimization problem by control the density of samples to refine the RBFNN.

Differential evolution (DE) was first introduced by Storn and Price [22] to solve the various global optimization problems. DE is a manageable yet powerful evolutionary algorithm with the advantages of less parameter, high simplicity, and fast convergence [22]. DE has been beneficial to various networks such as Hopfield Neural Network [23] and feed-forward neural networks [24]. Chauhan & Chandra [22] proposed the DE algorithm to train a wavelet neural network (WNN) by minimizing network error to obtain the proper relationship from the input vector in the input layer to the output vector in the output layer. Tao *et al.* [25] utilized the DE algorithm to improve RBFNN as the prediction model for the coking energy consumption process. Particle Swarm Optimization algorithm (PSO) is a nature-inspired evolutionary algorithm that imitates the influence of bird migration behavior [26]. PSO algorithm is one of the evolutionary algorithms proposed by Kennedy and Eberhart [27]. In some succeeding works, Qasem & Shamsuddin [28] proposed the PSO algorithm for enhancing RBFNN learning by optimizing the parameters of the hidden layer and output layer. Another study has been made by Alexandridis *et al.* [29], who used the PSO algorithm to optimize the construction of RBFNN. The proposed model was able to solve classification problems and solve function approximations with improved generalization capabilities and accuracy.

Karaboga and Basturk [30], [31] proposed the Artificial Bee Colony algorithm (ABC) to gain computational edge in optimizing the capability of both local search and global search. ABC was inspired by collective behaviors of bees gathering honey in an optimized pattern. ABC has been beneficial to various networks such as Hopfield Neural Network [14] and Hermite Neural Network [32]. Kurban & Besdok [33] utilized ABC to estimation the centers, width, and weights as the main parameters of RBFNN. Yu and Duan [34] proposed an optimized ABC in RBFNN integrated with Fuzzy C mean Clustering. In this paper, 2 layers of optimization in ABC were reported to increase the accuracy of the image fusion. Jafrasteh and Fathianpour [35] proposed hybrid RBFNN by introducing perturbation in ABC. The proposed system was reported to capture non-linear relationship in ore grade data. In another development, Satapathy *et al.* [36] combined the benefit of kernel trained ABC to further optimize the capability of RBFNN. The proposed RBFNN managed to increase the classification accuracy of EEG signal for epileptic seizure identification. The perspective has been expanded by Aljarah *et al.* [37] when they introduced hybrid ABC with RBFNN to solve well known datasets. On the perspective of logic programming in RBFNN, little studies have been done to optimize the parameter of RBFNN by using ABC. Kasihmuddin *et al.* [14] has demonstrated the ability of ABC to serve as an effective learning algorithm in Hopfield Neural Network (HNN). One of the notable use of ABC is proposed by Jiang *et al.* [30]. In this work, the ABC is employed for optimizing the parameters of RBFNN and predicting the ecological pressure. In another development, Menad *et al.* [38] have utilized the RBFNN framework with ABC algorithm (RBFNN-ABC) for predicting the carbon dioxide solubility and concentration in brine. The results manifested the capability of ABC in optimizing RBFNN that result in higher accuracy. By hybridizing RBFNN with 2SAT logic, here we examine the effects of ABC on the training phase as a single framework, RBFNN-2SATABC. Worth noting that the proposed model will be compared with the existing models. Thus, the main motivation of employing ABC in this research is due to:

1. According to Kasihmuddin *et al.* [14], [62], ABC has outperformed the other algorithm such as [5] and [6] in enhancing the training phase for bipolar 2SAT logical representation. We extended the non-binary representation for optimizing the parameter entrenched in the hidden layer of RBFNN as inspired by the binary operators consist of employed bees and onlooker bees' phase.

2. Several current studies such as Menad *et al.* [38] and Jiang *et al.* [39] utilize the ABC in optimizing the prediction capability of RBFNN. Both local search and global search capability reduce the chances for ABC to achieve sub-optimal fitness. Motivated by these recent works, ABC algorithm is applied in improving the output quality from the output weight thereby improving the performance of the structure RBFNN-2SAT.

To this end, the contributions of this paper are as follows:

1. This paper explores another perspective in approaching implicit knowledge by using an explicit learning model. Real-life problem (implicit representation) is learnable by using a set of explicit mathematical representation (2SAT logical rule).
2. This is the first attempt to embed 2SAT logical rule (knowledge) to the feed-forward neural networks (learner). In this study, the 2SAT logical rule has been embedded in RBFNN by systematically obtaining the optimal value of parameters (center and width). 2SAT logical rule is expected to optimize the structure of the RBFNN by fixing the number of hidden neurons involved.
3. Since the training of the proposed RBFNN always converges to suboptimal output weight, this paper will explore the capability of Artificial Bee Colony (ABC) compared to other existing established metaheuristics. The aim of the training model in RBFNN is to obtain the optimal output weight with the lowest iteration error. Extensive experimentation with various performance metrics has been conducted to reveal the effectiveness of ABC in the proposed RBFNN-2SAT.
4. The proposed RBFNN provides an interesting perspective. RBFNN obtained the output weight of 2SAT by minimizing the

objective function with the structurally systematic parameters. This approach is interestingly different from Sathasivam [40] that utilized the Wan Abdullah method in finding the correct synaptic weight (output weight). Although both paradigms utilized ABC in optimizing the proposed methods, the method proposed in this paper deals with non-binary optimization compared to the existing method. Therefore, the proposed method creates a new possible horizon for logic programming in the neural networks.

The rest of this paper is arranged as follows. The 2SAT logical rule is formulated in the first section. After the overview structure of the general RBFNN, the proposed hybrid model integrated with 2SAT is constructed. Accordingly, the proposed training model via metaheuristics algorithm namely GA, DE, PSO, and ABC will be discussed in detail. Finally, this paper presents numerical results to show the effectiveness of ABC in optimizing 2SAT in RBFNN and we conclude the paper with some remarks and future work.

## II. Boolean 2 Satisfiability Representation

Satisfiability (SAT) is demarcated as a logic rule with an array of clauses composed of binary literals. SAT is effectively governed by positive [5] and negative outcomes. The main structure of SAT representation is shown as follows:

(a) Consists of a set of $m$ variables of $v_1, v_2, v_3, \ldots, v_m$.

(b) Composes of a set of literals. A literal refers to the variable $v$ or a negation of a variable, $\neg v$.

(c) A set of $n$ discrete clauses, $l_1, l_2, l_3, \ldots, l_n$. Every single clause composes of literals strictly combined by only $\wedge$ logical operator.

Every variable can only take a bipolar value which is 1 or 0 that exemplifies the idea of true and false. Another variant of SAT representation is 2 Satisfiability. 2 Satisfiability (2SAT) consist of set of clauses that contain strictly 2 literals. The general formula for 2SAT logic is as follows:

$$P_{2SAT} = \bigwedge_{i=1}^{n} l_i, \text{where } l_i = \bigvee_{i=1}^{k} C_i \bigvee_{j=1}^{n} D_j, k = 2 \tag{1}$$

where $l_i$ refers to the clauses of 2SAT, meanwhile $C_i$ and $D_i$ denote the literals, $\vee$ refers to Disjunction (OR), and $\wedge$ is an logical operator of Conjunction (AND).

The goal of 2SAT logic is to establish the ideal logical model of RBFNN to calculate the parameters of the hidden layer which contribute in deciding the number of hidden neurons in the hidden layer. Ideally, a combinatorial problem is similar to an ordinary mathematical model with quantifiable rate of change. Unfortunately, that statement does not hold if the specific combinatorial problem is dynamical and appeared as non-linear or linearly distributed. There were several efforts to represent the combinatorial problem via 2SAT formulation [42], [43]. These combinatorial problems contain implicit knowledge and could not be represented in standard rate of changes [44]. From that perspective, 2SAT is the main representation because this logical rule has a huge flexibility in terms of state (1 or 0) compared to standard mathematical representation.

## III. Radial Basis Function Neural Network

Radial Basis Function Neural Network (RBFNN) is a variant of feed forward neural network with hidden interconnected layer which was pioneered by Lowe and Moody [45], [46]. Compared to other network, RBFNN has a more integrated structure and architecture. In terms of structure, RBFNN contains three neuron layers for computation purposes (See Fig. 1) [47]. In the input layer, $m$ neurons represent the input data that was transferred to the system. During the training phase, the parameters (center and width) will be calculated in the hidden layer. The parameters obtained will be used to calculate the output weight in the output layer. To reduce the dimensionality from the input to the output layer, a Gaussian activation function has been introduced. The Gaussian activation function, $\varphi_i(x)$ of the hidden neuron in RBFNN is as follows [48], [49]:

$$Q(x) = \frac{\left\| \sum_{j=1}^{N} w_{ji}' x_j - c_j \right\|^2}{2\sigma_j^2} \tag{2}$$

$$\varphi_i(x) = e^{-Q(x)} \tag{3}$$

where $c_j$, $\sigma_i$ are the center and width of the hidden neuron, respectively. In this case, $x_j$ is a input value for $N$ input neurons and the Euclidean norm $\| \|$ from neuron $i$ to $j$ can be defined as follows:

$$\left\| \sum_{j=1}^{N} w_{ji}' x_j - c_i \right\| = \sqrt{\sum_{m=1}^{N} \left( \sum_{j=1}^{N} w_{ji}' x_j - c_i \right)^2} \tag{4}$$

where $w_{ji}'$ is the input weight between the input neuron $j$ and the hidden neuron $i$. Structurally, $x_j$ is a input data in the training set and the hidden neuron $i$. $c_i$ is the center of the hidden neuron. The final output of RBFNN $F(w_i)$ is given by the following:

$$F(w_i) = \sum_{i=1}^{j} w_i \varphi_i(x_k) \tag{5}$$

where $F(w_i) = (F(w_1), F(w_2), F(w_3), \ldots, F(w_N))$ is the output value of RBFNN and the output weight is given by $w_i = (w_1, w_2, \ldots, w_N)$.

The aim of RBFNN is to obtain the optimal weights $w_i$ that satisfy the desired output value. In RBFNN, the hidden neuron provides a set of function that represents input pattern spanned by the hidden neuron [4], [47].

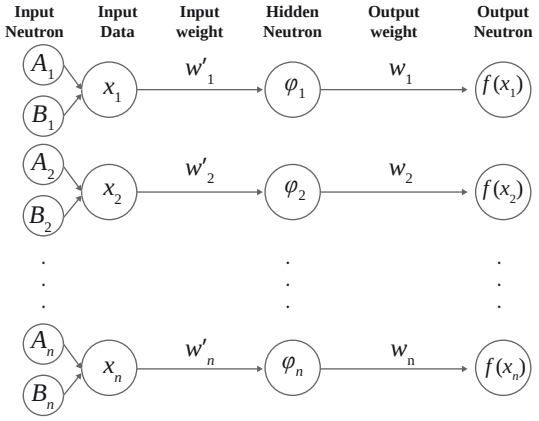



Fig. 1. Structure of RBFNN.

In this section, we will consider no training in conventional method Radial Basis Function Neural Network. Radial Basis Function Neural Network no-training paradigm was proposed by Vakil-Baghmisheh and Pavešić [50]. No training in Radial Basis Function is the simplest training because all the parameters were fixed. This method of training of RBFNN-2SAT does not have any practical value, because the number of prototype vectors should be equal to the number of

input data, and consequently the network will be too complex. Fig. 2 shows the steps to integrate RBFNN no training with 2SAT, which can be abbreviated as RBFNN-2SATNT [9]:

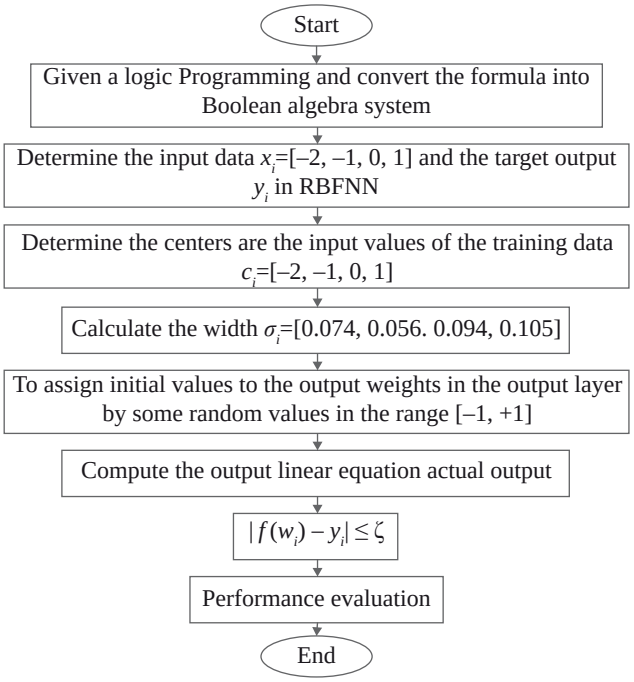



Fig. 2. Flowchart of RBFNN-2SATNT.

The parameter $x_i$ is the input data, whereas $c_i = x_i$ is the center, $\sigma_i$ is the width, and $\zeta$ is the tolerance value.

## IV. 2SAT Programming in RBFNN

Kasihmuddin *et al.* proposed logic programming by integrating 2SAT rule with neural network [14], [51]. The weight of the network was determined by Wan Abdullah method [2] where the inconsistencies of 2 Satisfiability logical rule have been minimized. The only problem of the proposed network is the rigidness of the weight calculation. 2SAT can be embedded to RBFNN by representing the variable as input neuron. Each input neuron $x_j$ constitutes {0,1} which signifies False and True. By using the value from input neuron, the parameters such as $c_i$ and $\sigma_i$ will be computed and the best number of hidden neuron will be obtained. In other words, embedding 2SAT as a logical rule makes RBFNN able to receive more input data with a fixed value of center and width. Hence the aim of the combination is to create a RBFNN model that classifies data based on 2SAT logical rule. Representation of 2SAT in RBFNN is given as the following formula:

$$P_{2SAT} = \bigvee_{i=1}^{k} C_i \bigvee_{j=1}^{n} D_j \tag{6}$$

where $k, n \in \mathbb{N}$. $C_i$ and $D_j$ are atoms. Applying embedding method of RBFNN, Eq. (6) will transform to:

$$x_i = \sum_{i=1}^{k} I(C_i) + \sum_{j=1}^{n} I(D_j) \tag{7}$$

$$I(C_i) \text{ or } I(D_j) = \begin{cases} 1, & \text{when } C \text{ or } D \text{ is True} \\ 0, & \text{when } C \text{ or } D \text{ is False} \end{cases} \tag{8}$$

Eq. (7) and (8) are vital in calculating training data for each 2SAT clause. Hence the implementation of 2SAT in RBFNN is abbreviated as RBFNN-2SAT. Table I illustrates the input data of RBFNN-2SAT for:

$$P_{2SAT} = C, D \leftarrow, E \leftarrow F, K \leftarrow L \tag{9}$$

TABLE I. The Input Data and the Output Target data for $P_{2SAT} = C, D \leftarrow, E \leftarrow F, K \leftarrow L$

| Clause | $C, D \leftarrow$ | | | $E \leftarrow F$ | | | $K \leftarrow L$ | | |
|---|---|---|---|---|---|---|---|---|---|
| DNF | $C \vee D$ | | | $E \vee \neg F$ | | | $K \vee \neg L$ | | |
| The Input Data Form | $x = C + D$ | | | $x = E - F$ | | | $x = K - L$ | | |
| Input Data in the Training Set $x_i$ | 0 | 1 | 2 | -1 | 0 | 1 | -1 | 0 | 1 |
| The Target Output Data $y_i$ | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |

After finding the center and the width of the hidden layer, RBFNN will use the Gaussian function in Eq. (3) to calculate the output weight. As the number of clauses increase, RBFNN-2SAT requires more efficient learning method to find the correct output weight. In this paper, a metaheuristics algorithm will be implemented to find the optimal output weights that minimize the following objective function:

$$f(w_i) = \sum_{i=1}^{j} w_i \varphi_i(x) \tag{10}$$

where $f(w_i)$ is the final output classification of the RBFNN-2SAT.

## V. Genetic Algorithm in RBFNN-2SAT

A Genetic Algorithm (GA) is a standard metaheuristic algorithm in solving various optimization problems. Given a finite solution space, the structure of a GA can be divided into local search and global search [52]. In a GA, the strings populations called chromosomes are represented in terms of solutions to the optimization problem [53]. The quality of the chromosome is denoted by the fitness value. At every generation, the fitness value of each chromosome is estimated, and the best fitness is selected as final solution. The chromosomes improve their fitness by implementing three (3) operators namely crossover, selection and mutation. Crossover promotes the exchange of information between chromosomes. Hamadneh *et al.* [4] used the GA to decide the centers of hidden neurons width and number of the hidden neuron by minimize the sum of absolute error of the actual outputs and the desired outputs. During selection, several chromosomes are selected from the current population depending on their fitness value. Mutation has been added to create genetic diversity of the chromosomes. In this paper, GA will be used to optimize the output weight of RBFNN-2SAT by reducing the training error. The implementation of GA in RBFNN is defined as RBFNN-2SATGA. In RBFNN-2SATGA, GA will calculate the output weight by using the centers, width in the hidden neuron. The steps involved in RBFNN-2SATGA are as follows:

**Step 1**

**Population Initialization**: The output weights represented by a chromosome will be initialized. The representations of chromosomes are as follows:

$$w_i = (w_1, w_2, w_3, ...., w_N) \tag{11}$$

The population has $N_{pop}$ chromosomes containing $N_N$ of random output weights. The aim is to minimize the objective function:

$$f_{GA}(w_i)=\begin{cases}1, & \sum_{i=1}^{j} w_i \varphi_i(x) \le 0\\ 0, & Otherwise\end{cases} \quad (12)$$

where $f_{GA}(w_i)$ is the objective function in the RBFNN-2SATGA model.

**Step 2**

**Fitness Computation:** The fitness of each individual chromosome is calculated via a basis function of RBFNN-2SAT. The basis function used in this paper is shown in the following equation:

$$fit_i = \frac{1}{\left(1+f_{GA}(w_i)\right)},\ 0 \le fit_i \le 1 \quad (13)$$

where $f_{GA}(w_i)$ is the objective function and $fit_i$ is the fitness value.

**Step 3**

**Selection:** The chromosomes are arranged in descending order based on the value of the fitness function. Only the best chromosomes (with the highest fitness value) are kept while others are discarded. The selection probability, $p_i$ for each chromosome will be calculated by using the following equation:

$$p_i = \frac{fit_i}{\sum_{i=0}^{n} fit_i} \quad (14)$$

**Step 4**

**Crossover:** During the crossover phase, information from the parent will be randomly exchanged for creating offspring with different genetic composition. The location of the crossover will be randomly selected. Crossover phase will determine the number of cross-population according to the crossover rate. Given two parents $w_k$ and $w_m$, the offspring $w_i^{new}$ will be produced by the following equations [54], [55]:

$$w_i^{new}=\begin{cases}w_m + r(w_k - w_m), & p_i,\ i=1,2,3,...n\\ w_m, & 1-p_i\end{cases} \quad (15)$$

where $p_i$ is the probability, $r$ is the crossover rate, $w_k$ is the chromosome with higher probability, $w_m$ is the chromosome with lower probability and the parameter $k$ is choosen by the following equation:

$$k=\begin{cases}rand(m,n), & p_i\\ m, & 1-p_i\end{cases} \quad (16)$$

where $k + m = n$ and $k > m$. The value of $k$ is uniformly distributed between $k$ and $m$.

**Step 5**

**Mutation:** During the mutation phase, the chromosome information will be randomly assigned within the pre-determined range (often determined by the user). The mutation is expected to create a newly breed of chromosome. The equation involved is as follows:

$$w_m^{new}=\begin{cases}rand(-5,5), & rand(0,1)<\tau\\ w_i, & rand(0,1)\ge\tau\end{cases} \quad (17)$$

where $w_m^{new}$ is the new chromosome from mutation phase when $\tau \in [0,1]$.

**Step 6**

**Termination:** GA will iterate up to 10000 Generations. If a given solution termination criterion is met, the calculation of the algorithm is stopped or will go back to step 2 with $i = i+1$. The final output of RBFNN-2SAT is a chromosome that contains optimal output weights of RBFNN-2SATGA.

## VI. Differential Evolution Algorithm in RBFNN-2SAT

Storn and Price [22] has fruitfully introduced a new evolutionary population-based algorithm called the Differential Evolutionary (DE) algorithm which typically is being used in numerical optimization. The fundamental framework of DE algorithm can be divided into local and global search with an adaptable function optimizer [56]. The core differences between GA and DE is that the selection operator in DE uses an equal probability to elect parents. Hence, the chance is independent towards the fitness value of the solutions. In the DE algorithm, every individual solution competes with its parent and the fittest one will win [57]. In this work, the DE algorithm will be adopted as a learning mechanism during the training phase. The purpose of the training is to compute the corresponding output weights that connect hidden neurons and output neurons of RBFNN-2SAT. The stages involved in RBFNN-2SATDE in optimizing the connection weights between the hidden layer and the output layer is represented in Fig. 3.

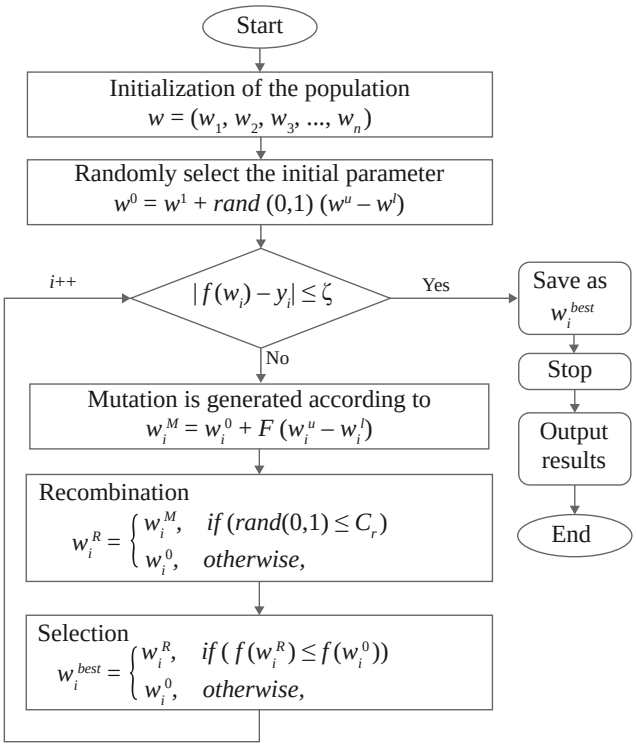



Fig. 3. Flowchart of RBFNN-2SATDE.

The real parameters $w^l$ and $w^u$ are lower and upper bounds, respectively. $w^0$ is the initial parameter value distributed uniformly on the intervals $[w_i^l, w_i^u]$. $w_i^M$ is the mutation output weight, $F \in [0,2]$ is the mutation factor. $w_i^R$ is the recombination output weight, $C_r \in [0,1]$ is the crossover probalitiy. $\zeta$ is the tolerance value.

## VII. Particle Swarm Optimization Algorithm in RBFNN-2SAT

The PSO algorithm is a class of iterative swarm-based searching

algorithm, deployed widely as the learning algorithm or universal optimization. The pioneer work of PSO was coined by Eberhart and Kennedy [26] by mathematically modelling the socio-behavioral feature of the bird flocking and fish schooling in their own population. The remarkable feature in PSO is the existence of adjustable free parameters, which makes it easy to implement and optimize. Specifically, PSO adopted a vigorous searching process by impending the best particle in a solution space [58]. Pursuing that, the potential solutions, named particles, fly over the searching space by succeeding the existing optimum particles. In addition, the changes in the position of the particles occur in PSO, where it is vital in searching for the best particle. This study adopts the PSO algorithm to optimize the output weight among the hidden neurons and the output neurons of RBFNN-2SAT. Therefore, the steps involved in RBFNN-2SATPSO are represented in Fig. 4.

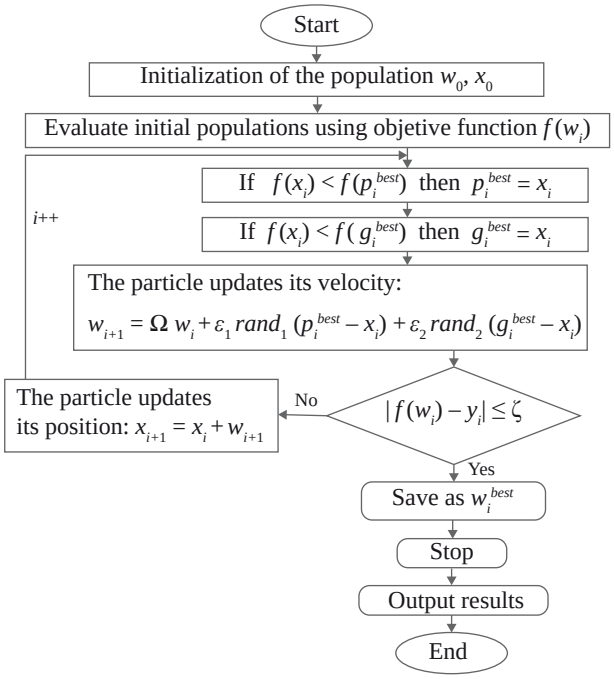



Fig. 4. Flowchart of RBFNN-2SATPSO.

The parameter $\Omega$ is the inertia weight, whereas $\varepsilon_1 = \varepsilon_2 = 2$ are acceleration constants, $rand_1 = rand_2$ are experimented arbitrarily within [0, 1], $p_i^{best}$ refers to the individual best position attained by the particle of the primary swarm, and $g_i^{best}$ denotes the global best position completed by the particles of the sucessive swarm and the position of the new particle, $x_i$. Additionally, $\zeta$ is the tolerance value.

## VIII. Artificial Bee Colony Algorithm in RBFNN-2SAT

Artificial bee colony (ABC) algorithm has been introduced by Karaboga [59] in resolving various mathematical optimization problems. In ABC, the colony of bees contains three groups called employed bees, onlooker bees, and scout bees. Generally, employed bees bring quantities of nectar from the resource food to the hive. They will share the information about the source of food with a certain probability by dancing inside the hive. Then, onlooker bees stay in the dancing areas and decide source of food depending on the prospect (the probability) provided by the employed bees [32]. The other type of bees is called the Scout Bee, which conducts the random search for new sources of food if the quality of the food source is not in a satisfactory state. In this paper, ABC will be used to optimize the output weight of RBFNN-2SAT by reducing the training error. The implementation of ABC in RBFNN is defined as RBFNN-2SATABC. In this context, the function to be optimized is:

$$f_{ABC}(w_i) = \begin{cases} 1, & \sum_{i=1}^{j} w_i\, \varphi_i(x) \le 0 \\ 0, & Otherwise \end{cases} \tag{18}$$

where $f_{ABC}(w_i)$ is the objective function of the RBFNN-2SATABC model. The algorithm involved in RBFNN-2SATABC is as follows:

Step 1

**Population Initialization**: Initialize all the bee that is:

$$w_{ji} = (w_{1,i}, w_{2,i}, ..., w_{ji}, ..., w_{di}) \tag{19}$$

in RBFNN-2SAT as:

$$w_{ji} = w_{j\min} + rand[0,1]\,(w_{j\max} - w_{j\min}) \tag{20}$$

where $w_{ji} \in \left[ w_{j\min}, w_{j\max} \right]$, $w_{j\min}$ and $w_{j\max}$ are the minimum value and maximum value of the output weight with index of $i \in \{1,2,...,n\}$ and $j \in \{1,2,...,d\}$. $n$ is the number of employed bees (the number of solutions), and $d$ is the dimension of the solution space (number of hidden neurons).

Step 2

**Employed Bee Phase:** Employed bee will search for the food source. The new food source (solution) for employed bees, $w_{ji}^{employed}$ is given as follows:

$$w_{ji}^{employed} = w_{ji} + rand\left[0,1\right]\left(w_{ji} - w_{jk}\right) \tag{21}$$

where $j$, $k$ are selected randomly and the $w_{jk}$ is called the neighbor bee of $w_{ji}$. The value of $f_{ABC}\left(w_{ji}^{employed}\right)$ will be calculated as follows:

$$f_{ABC}\left(w_{ji}^{employed}\right) = \sum_{i=1}^{j} w_{ji}^{employed}\, \varphi_i(x) \tag{22}$$

$$fit_i = \frac{1}{1 + f_{ABC}\left(w_{ji}^{employed}\right)} \tag{23}$$

where $fit_i$ is the fitness value of the bee.

Step 3

**Onlooker Bee Phase:** The probability value of the food sources will be calculated. Onlooker bee will perform exhange of information based on the following probability:

$$p_i^{\text{Onlooker}}{}_i = \frac{fit\left(w_i^{employed}\right)}{\sum_{i=1}^{SN} fit\left(w_i^{employed}\right)} \tag{24}$$

By using the above probability, the food source will be obtained by using equation (21).

Step 4

**Scout Bee Phase:** If the values of fitness of the employed bees are not improving by a number continuous predetermined of iterations, which is called (*Limit*) those food source are abandoned, and these employed bee become the scouts, and generate a new solution $w_i^{new}$ for the employed bee by using the following equation:

$$w_i^{new} = \begin{cases} rand(-5,5), & limit > trial \\ w_i, & Otherwise \end{cases} \quad (25)$$

Step 5

**Termination:** If the stopping criterion is met, then it stops and the best food source is memorized, otherwise, the algorithm returns to Step 2.

## IX. Experimental Setup

All the proposed RBFNN-2SAT model will be executed and coded in Microsoft Visual C # 2008 Express program in Microsoft Window 7, 64-bit, with 500 GB hard drive specification, 4096 MB RAM, and 3.40 GHz processor. The lists of parameters used in each RBFNN-2SAT model are summarized in Table II to Table V. Simulated data sets will be obtained by randomly generate the input data. The choice of data reduces the possible bias of the data which covers a wider range of search space. Next, the number of neurons *NN* used in the experiment varies from $6 \leq NN \leq 108$.

TABLE II. List of Parameters in RBFNN-2SATGA

| Parameter | Value |
|---|---|
| Number of iteration | 10000 |
| Selection type | Wheel selection |
| Number of individuals | 50 |
| Mutation ratio | 1 |
| Mutation type | Uniform |
| Crossover ratio | 1 |
| Crossover type | Single point |

TABLE III. List of Parameters in RBFNN-2SATDE

| Parameter | Value |
|---|---|
| Number of iteration | 10000 |
| $C_r$ | [0, 1] |
| $F$ | [0, 2] |
| Population | 50 |

TABLE IV. List of Parameters in RBFNN-2SATPSO

| Parameter | Value |
|---|---|
| Ω | 0.6 |
| $\varepsilon_1$ | 2 |
| $\varepsilon_2$ | 2 |
| $rand_1 = rand_2$ | [0,1] |
| Number of iteration | 10000 |

TABLE V. List of Parameters in RBFNN-2SATABC

| Parameter | Value |
|---|---|
| No_Employed_bees | 50 |
| No_Onlooker_bees | 50 |
| No_Scout_bees | 1 |
| *Limit* | 1000 |
| *Trial* | 10000 |

## X. Results and Discussion

Hamadneh *et al.* [60] use mean square error as a metric to appraise the performance of the trained RBFNN. In this paper, both proposed hybrid models will be compared by using four performance metrics such as Root Mean Square Error (RMSE), Sum of Squares Error (SSE), Mean Absolute Percentage Error (MAPE) and CPU Time. The equation for each performance metrics is as follows:

$$RMSE = \sum_{i=1}^{n} \sqrt{\frac{1}{n}\left(f(w_i) - y_i\right)^2} \quad (26)$$

$$SSE = \sum_{i=1}^{n} \left(f(w_i) - y_i\right)^2 \quad (27)$$

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{\left(f(w_i) - y_i\right)}{y_i} \right| \quad (28)$$

where $f(w_i)$ is the actual output value, $y_i$ is the target output value and $n$ is number of the iterations. In addition, computation time will be considered in order to evaluate the efficiency of the RBFNN model.

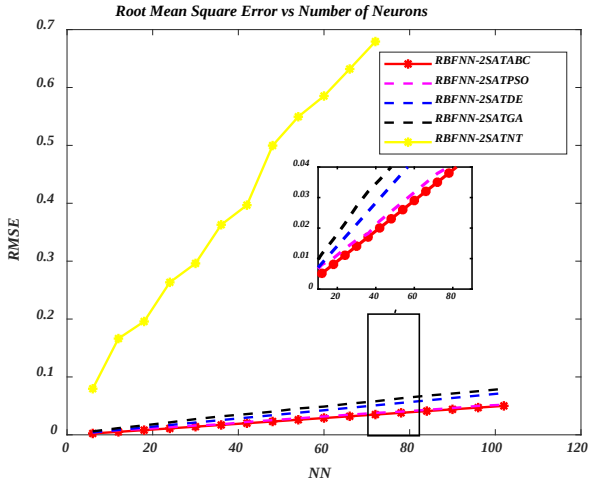



Fig. 5. RMSE value for all RBFNN-2SAT models.

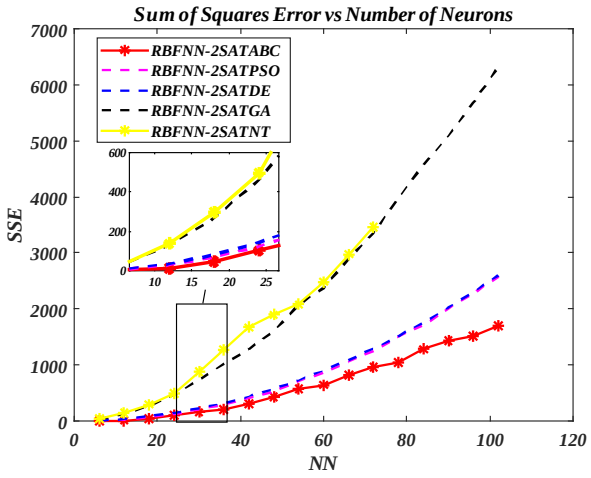



Fig. 6. SSE evaluation.

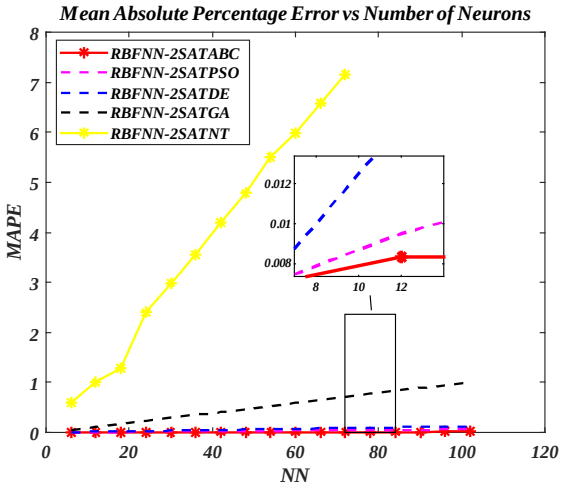



Fig. 7. MAPE evaluation.

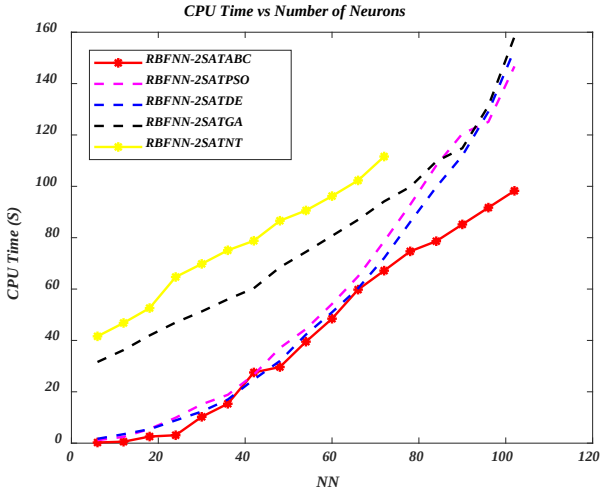



Fig. 8. Computation time evaluation.

In this study, 2SAT logical rule is expected to perform comparatively exceptional to other non-systematic logical rule such as [6], [29], [61], [62], [63]. This is due to the variation of the number of variables in each clause. This causes RBFNN-2SAT to alter the dimension of the hidden layer. Imbalance signal from the hidden layer to the output layer will lead to imbalance value of parameters (centre and width) and high computation error. The results in Fig. 5 until Fig. 8 allow to deduce the following findings:

1. RBFNN-2SAT can receive more input data with a fixed value of center and width. In this case, RBFNN-2SATABC creates a model that classifies data based on 2SAT logical rule with minimum value of RMSE, SSE and MAPE.
2. RBFNN-2SATABC has best performances in terms of errors as the number of neurons is increased. In the exploration front (employed bee), ABC locates the general range of the optimal output weight. The value of the output weight improves significantly during the exploitation phase (onlooker bees). Based on the result, the probability for RBFNN-2SATABC to reach the scout bee phase is approximately zero. In this case, RBFNN-2SATABC effectively explores different solution space in less iterations.
3. In terms of computation time, RBFNN-2SATABC was reported to be faster than the other RBFNN-2SAT model. At $NN > 20$, the possibility for the conventional method RBFNN-2SATNT to be trapped in trial and error state increases. Trial and error cause RBFNN-2SATNT to achieve pre-mature convergence.
4. On the other hand, RBFNN-2SATGA has a relatively larger learning error because of ineffective initial crossover. It requires several iterations for RBFNN-2SATGA to produce high quality output weight. During that time, the only operator that is effective is mutation. The problem is worsened when the suboptimal output weight is a floating number.
5. RBFNN-2SATDE is reported to illustrate some drawbacks such as tendency to be trapped at sub-optimal output weight and slow convergence rate. In this case, RBFNN-2SATDE requires more iterations to satisfy $\left| f\left(w_i\right) - y_i \right| \le \zeta$ which results in the accumulation of error. In addition, the unbounded mutation operator in DE tends to create numerous alternate search space that reduces the probability of the RBFNN-2SATDE to achieve optimal output weight.
6. In another perspective, RBFNN-2SATPSO has a relatively lower learning error compared to another model. This is due to the use of the particle in this algorithm that mimics our proposed ABC algorithm. Although the result for RBFNN-2SATPSO seems quite promising, this algorithm lacks the control of the effective local search. In this case, as $t \to 10000$, the search space for each particle will magnify indefinitely and result in suboptimal output weight. Hence, RBFNN-2SATPSO will converge prematurely.

These experiments show that the ABC algorithm can be successfully applied to train RBFNN-2SAT. Another observation is that the effectiveness of ABC can be seen vividly when the number of neurons increases. Moreover, ABC algorithm in RBFNN achieves more promising performance based on RSME by 94.8%, SSE by 72.9%, MAPE by 99.1%, and CPU time by 39.8%. This concludes that ABC in RBFNN-2SAT could be used in practice to achieve better prediction results for the 2SAT logic programming.

## XI. Conclusion

A hybrid paradigm, ABC algorithm incorporated with RBFNN and 2SAT (RBFNN-2SATABC) has been fruitfully developed to foster the learning phase with different number of neurons. Following that, the work as reported in this paper reveals the significant differences in the performance of RBFNN-2SATABC in terms of Root Mean Square Error (RMSE), Sum of Squares Error (SSE), Mean Absolute Percentage Error (MAPE), and process time (computation time in seconds). Furthermore, the proposed paradigm offers an error of approximately 2% of MAPE, and faster computation time compared to RBFNN-2SATGA. Henceforth, the RBFNN-2SATABC has been clearly recognized to be more robust than the RBFNN-2SATGA in certain aspects which include better lower error and faster process time in performing 2SAT logic programming. As future development, the RBFNN-2SATABC can be improved by using different classes of Satisfiability logic ranging from, Major Satisfiability (MAJ-SAT), Weighted SAT, Maximum Satisfiability (MAX-SAT) and Unsatisfiable Satisfiability (MIN-UNSAT). This work also can be applied as a traditional optimization method to solve problems such as travelling salesman and N-queen's problem.

## Acknowledgment

This research was supported by Fundamental Research Grant Scheme (FRGS), Ministry of Education Malaysia, grant number 203/PMATH/6711804 and Universiti Sains Malaysia (USM).

## References

[1] M. S. Alkhaawneh, (2019). Hybrid Cascade Forward Neural Network with Elman Neural Network for Disease Prediction. *Arabian Journal for Science and Engineering*, 44(11), 9209-9220.

[2] W. A. T. W. Abdullah, (1992). Logic programming on a neural network. *International journal of intelligent systems*, 7(6), 513-519.

[3] S. Sathasivam, (2010). Upgrading logic programming in Hopfield network. *Sains Malaysiana*, 39(1), 115-118.

[4] N. Hamadneh, , S. Sathasivam, S. L. Tilahun, & O. H. Choon, (2012). Learning logic programming in radial basis function network via genetic algorithm. *Journal of Applied Sciences(Faisalabad),* 12(9): 840-847.

[5] M. S. M. Kasihmuddin, M. A. Mansor, & S. Sathasivam, (2017). Hybrid Genetic Algorithm in the Hopfield Network for Logic Satisfiability Problem. *Pertanika Journal of Science & Technology*, 25(1), 139 - 152.

[6] M.A.B. Mansor, M.S.B.M. Kasihmuddin, and S. Sathasivam, 2017. Robust Artificial Immune System in the Hopfield network for Maximum k-Satisfiability. *International Journal of Interactive Multimedia and Artificial Intelligence*, 4(4), 63-71.

[7] N. Hamadneh, W. A. Khan, I. Khan, & A. S. Alsagri, (2019). Modeling and Optimization of Gaseous Thermal Slip Flow in Rectangular Microducts Using a Particle Swarm Optimization Algorithm. *Symmetry*, 11(4), 488-491.

[8] H. de Leon-Delgado, R. J. Praga-Alejo, D. S. Gonzalez-Gonzalez, & M. Cantú-Sifuentes, (2018). Multivariate statistical inference in a radial basis function neural network. *Expert Systems with Applications*, 93, 313-321.

[9] S. Alzaeemi, M.A. Mansor, M.S.M. Kasihmuddin, S. Sathasivam, and M. Mamat, (2020). Radial basis function neural network for 2 satisfiability programming. *Indonesian Journal of Electrical Engineering and Computer Science*, 18(1), 459-469.

[10] M. H. Horng, Y. X. Lee, M. C. Lee, & R. J. Liou, (2012). Firefly meta-heuristic algorithm for training the radial basis function network for data classification and disease diagnosis. In Theory and new applications of swarm intelligence. *IntechOpen*, 10(19), 7-28.

[11] J. Yang, & J. Ma, (2019). Feed-forward neural network training using sparse representation. *Expert Systems with Applications*, 116, 255-264.

[12] M. S. M. Kasihmuddin, M. A. Mansor, M. B. M. Faisal, & S. Sathasivam, (2019). Discrete Mutation Hopfield Neural Network in Propositional Satisfiability. *Mathematics*, 7(11), 1133-1154.

[13] M.S.M., Kasihmuddin, Mansor, M.A. and Sathasivam, S., 2018. Discrete Hopfield Neural Network in Restricted Maximum k-Satisfiability Logic Programming. *Sains Malaysiana*, 47(6), 1327-1335.

[14] M. S. M. Kasihmuddin, M. A. Mansor, & S. Sathasivam, (2017). Robust Artificial Bee Colony in the Hopfield Network for 2-Satisfiability Problem. *Pertanika Journal of Science & Technology*, 25(2), 453 - 468.

[15] N. Hamadneh, S. Sathasivam, and O.H. Choon, (2012). Higher order logic programming in radial basis function neural network. *Appl Math Sci*, 6(3), 115-127.

[16] H. V. H. Ayala, & L.dos Santos Coelho, (2016). Cascaded evolutionary algorithm for nonlinear system identification based on correlation functions and radial basis functions neural networks. *Mechanical Systems and Signal Processing,* 1(68), 378-393.

[17] R. D. Dandagwhal, & V. D. Kalyankar, (2019). Design Optimization of Rolling Element Bearings Using Advanced Optimization Technique. *Arabian Journal for Science and Engineering*, 44(9), 7407-7422.

[18] J. H. Holland, (1973). Genetic algorithms and the optimal allocation of trials. *SIAM Journal on Computing*, 2(2), 88-105.

[19] D. E. Goldberg, & J. H. Holland, (1988). Genetic algorithms and machine learning, *Machine Learning*, 2(3), 95-99.

[20] M. Pandey, M. Zakwan, P. K. Sharma, & Z. Ahmad, (2020). Multiple linear regression and genetic algorithm approaches to predict temporal scour depth near circular pier in non-cohesive sediment. *ISH Journal of Hydraulic Engineering*, 26(1), 96-103.

[21] Z. Jing, J. Chen, & X. Li, (2019). RBF-GA: An adaptive radial basis function metamodeling with genetic algorithm for structural reliability analysis. *Reliability Engineering & System Safety*, 189, 42-57.

[22] R. Storn and K. Price, (1997). Differential evolution a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4), 341-359.

[23] A. Saha, A. Konar, P. Rakshit, A. L. Ralescu, & A. K. Nagar, (2013, August). Olfaction recognition by EEG analysis using differential evolution induced Hopfield neural net. *In The 2013 International Joint Conference on Neural Networks,* 4(9), 1-8.

[24] J. Ilonen, J. K. Kamarainen, & J. Lampinen, (2003). Differential evolution training algorithm for feed-forward neural networks. *Neural Processing Letters*, 17(1), 93-105.

[25] W. Tao, J. Chen, Y. Gui, & P. Kong, (2019). Coking energy consumption radial basis function prediction model improved by differential evolution algorithm. *Measurement and Control*, 52(8), 1122-1130.

[26] R. Eberhart, & J. Kennedy, (1995, October). A new optimizer using particle swarm theory. *In MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science,* 39-43.

[27] J. Kennedy, & R. Eberhart, (1995, November). Particle swarm optimization. In Proceedings of ICNN'95-*International Conference on Neural Networks,* vol. 4, 1942-1948.

[28] S. N. Qasem, & S. M. H. Shamsuddin, (2009, May). Improving performance of radial basis function network based with particle swarm optimization. *In 2009 IEEE Congress on Evolutionary Computation, Man and Cybernetics*, 3149-3156.

[29] A. Alexandridis, E. Chondrodima, & H. Sarimveis, (2016). Cooperative learning for radial basis function networks using particle swarm optimization. *Applied Soft Computing,* 49, 485-497.

[30] D. Karaboga, & B. Basturk, (2007). A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. *Journal of global optimization*, 39(3), 459-471.

[31] D. Karaboga, & E. Kaya, (2019). Training ANFIS by using an adaptive and hybrid artificial bee colony algorithm (aABC) for the identification of nonlinear static systems. *Arabian Journal for Science and Engineering*, 44(4), 3531-3547.

[32] G.E. Tsekouras, V. Trygonis, A. Maniatopoulos, A. Rigos, A. Chatzipavlis, J. Tsimikas, N. Mitianoudis, and A.F. Velegrakis, (2018). A Hermite neural network incorporating artificial bee colony optimization to model shoreline realignment at a reef-fronted beach. *Neurocomputing*, 280, 32-45.

[33] T. Kurban, & E. Beşdok, (2009). A comparison of RBF neural network training algorithms for inertial sensor based terrain classification. *Sensors*, 9(8), 6312-6329.

[34] J. Yu, & H. Duan, (2013). Artificial bee colony approach to information granulation-based fuzzy radial basis function neural networks for image fusion. *Optik-International Journal for Light and Electron Optics*, 124(17), 3103-3111.

[35] B. Jafrasteh, & N. Fathianpour, (2017). A hybrid simultaneous perturbation artificial bee colony and back-propagation algorithm for training a local linear radial basis neural network on ore grade estimation. *Neurocomputing*, 235, 217-227.

[36] S. K. Satapathy, S. Dehuri, & A. K. Jagadev, (2017). ABC optimized RBF network for classification of EEG signal for epileptic seizure identification. *Egyptian Informatics Journal,* 18(1), 55-66.

[37] I. Aljarah, H. Faris, S. Mirjalili, & N. Al-Madi (2018). Training radial basis function networks using biogeography-based optimizer. *Neural Computing and Applications*, 29(7), 529-553.

[38] N. A. Menad, A. Hemmati-Sarapardeh, A. Varamesh, & S. Shamshirband, (2019). Predicting solubility of CO2 in brine by advanced machine learning systems: Application to carbon capture and sequestration. *Journal of CO2 Utilization*, 33, 83-95.

[39] S. Jiang, C. Lu, S. Zhang, X. Lu, S. B. Tsai, C. K. Wang, & C. H. Lee, (2019). Prediction of Ecological Pressure on Resource-Based Cities Based on an RBF Neural Network Optimized by an Improved *ABC Algorithm. IEEE Access*, 7, 47423-47436.

[40] S. Sathasivam, (2010). Upgrading logic programming in Hopfield network. *Sains Malaysiana*, 39(1), 115-118.

[41] T. Hoeink, (2019). Boolean satisfiability problem for discrete fracture network connectivity. *Patent Application Publication*, 180(53), 1-12.

[42] S. Mukherjee, & S. Roy, (2015). Multi terminal net routing for island style FPGAs using nearly-2-SAT computation. In VLSI Design and Test (VDAT), 2015 *19th International Symposium on IEEE*, 10(1109), 1-6.

[43] R. Miyashiro, & T. Matsui, (2005). A polynomial-time algorithm to find an equitable home away assignment. *Operations Research Letters*, 33(3), 235-241.

[44] S. Even, A. Itai, & A. Shamir, (1976). On the Complexity of Timetable and Multicommodity Flow Problems. *SIAM Journal on Computing*, 5(4), 691-703.

[45] J. Moody, & C. J. Darken, (1989). Fast learning in networks of locally-tuned processing units. *Neural computation*, 1(2), 281-294.

[46] D. Lowe, (1989, October). Adaptive radial basis function nonlinearities,

and the problem of generalisation. *In Artificial Neural Networks, First IEE International Conference,* 1(313), 171-175.

[47] A. K. Hassan, M. Moinuddin, U. M. Al-Saggaf, & M. S. Shaikh, (2018). On the kernel optimization of radial basis function using nelder mead simplex. *Arabian Journal for Science and Engineering*, 43(6), 2805-2816.

[48] A. Idri, A. Zakrani, & A. Zahi, (2010). Design of radial basis function neural networks for software effort estimation. *IJCSI International Journal of Computer Science Issues*, 7(4), 11-17.

[49] S. B. Roh, S. K. Oh, W. Pedrycz, K. Seo, & Z. Fu, (2019). Design methodology for Radial Basis Function Neural Networks classifier based on locally linear reconstruction and Conditional Fuzzy C-Means clustering. *International Journal of Approximate Reasoning*, 106, 228-243.

[50] M. T. Vakil-Baghmisheh, & N. Pavešić, (2004). Training RBF networks with selective backpropagation. *Neurocomputing*, 62, 39-64.

[51] L.C. Kho, M. S. M. Kasihmuddin, M. A. Mansor, & S. Sathasivam, (2020). Logic Mining in League of Legends. *Pertanika Journal of Science & Technology*, 28(1), 211 - 225.

[52] W. Jia, D. Zhao, T. Shen, C. Su, C. Hu, & Y. Zhao, (2014). A new optimized GA-RBF neural network algorithm. *Computational intelligence and neuroscience*, 1(4), 1-6.

[53] H. Marouani, K. Hergli, H. Dhahri, & Y. Fouad, (2019). Implementation and Identification of Preisach Parameters: Comparison Between Genetic Algorithm, Particle Swarm Optimization, and Levenberg–Marquardt Algorithm. *Arabian Journal for Science and Engineering*, 44(8), 6941-6949.

[54] M. Awad, (2010). Optimization RBFNNs parameters using genetic algorithms: applied on function approximation. *International Journal of Computer Science and Security (IJCSS),* 4(3), 295-307.

[55] L. J. Eshelman, & J. D. Schaffer, (1993). Real-coded genetic algorithms and interval-schemata. *In Foundations of genetic algorithms*, Vol. 2. Elsevier, 187-202.

[56] S. L. Wang, F., Ng, T. F. Morsidi, H. Budiman, & S. C. Neoh, (2020). Insights into the effects of control parameters and mutation strategy on self-adaptive ensemble-based differential evolution. *Information Sciences*, 514, 203-233.

[57] K. R. Opara, & J. Arabas, (2019). Differential Evolution: A survey of theoretical analyses. *Swarm and evolutionary computation*, 44, 546-558.

[58] Y. Fukuyama, & H. Yoshida, (2001, May). A particle swarm optimization for reactive power and voltage control in electric power systems. *In Proceedings of the 2001 Congress on Evolutionary Computation* (IEEE Cat. No. 01TH8546), vol. 1, 87-93.

[59] D. Karaboga, (2005). An idea based on honey bee swarm for numerical optimization. *Technical reporttr 06, Erciyes university, engineering faculty, computer engineering department*, Vol. 200, 1-10.

[60] N. Hamadneh, S. Sathasivam, S. L. Tilahun, & O. H. Choon, (2014, July). Satisfiability of logic programming based on radial basis function neural networks. *In AIP Conference Proceedings,* 1605(1), 547-550.

[61] M. S. M. Kasihmuddin, M. A. Mansor, & S. Sathasivam, (2016). Genetic Algorithm for Restricted Maximum k-Satisfiability in the Hopfield Network. *International Journal of Interactive Multimedia & Artificial Intelligence*, 4(2), 52-60.

[62] M. S. M. Kasihmuddin, M. A. Mansor, & S. Sathasivam, (2016). Artificial Bee Colony in the Hopfield Network for Maximum k-Satisfiability Problem. *Journal of Informatics and Mathematical Sciences*, 8(5), 317-334.

[63] C. Caleiro, F. Casal, & A. Mordido, (2019). Generalized probabilistic satisfiability and applications to modelling attackers with side-channel capabilities, *Theoretical Computer Science,* 781, 39-62.

**Mohd Shareduwan Mohd Kasihmuddin**

Mohd Shareduwan Mohd Kasihmuddin is a lecturer in School of Mathematical Sciences, Universiti Sains Malaysia. He received his Ph.D from Universiti Sains Malaysia. His current research interests include Metaheuristics method, neural network development, artificial intelligence and logic programming. He can be contacted via shareduwan@usm.my.

**Mohd. Asyraf Mansor**

Mohd. Asyraf Mansor is a lecturer in School of Distance Education, Universiti Sains Malaysia. He received his Ph.D from Universiti Sains Malaysia. His current research interests include evolutionary algorithm, satisfiability problem, neural networks, logic programming and heuristic method.

**Shehab Abdulhabib Alzaeemi**

Shehab Abdulhabib Alzaeemi received a Bachelor Degree of Education (Science) from Taiz Universiti in 2004, Master of Science (Mathematics) from Universiti Sains Malaysia in 2016 and an ongoing PhD student in Universiti Sains Malaysia. He was a fellow under the Academic Staff Training System of Sana'a Community College from 2005-2014. His research interests mainly focus on neural network, logic programming, and data mining. His email is shehab_alzaeemi@yahoo.com

**Saratha Sathasivam**

Saratha Sathasivam is an Associate Professor in the School of Mathematical Sciences, Universiti Sains Malaysia. She received her MSc and BSc(Ed) from Universiti Sains Malaysia. She received her Ph.D at Universiti Malaya, Malaysia. Her current research interest are neural networks, agent based modeling and constrained optimization problem. Her email is saratha@usm.my.

# A Hybrid Approach for Android Malware Detection and Family Classification

Meghna Dhalaria, Ekta Gandotra*

Department of Computer Science and Engineering, Jaypee University of Information Technology, Waknaghat, Solan, HP (India)

Received 13 February 2020 | Accepted 29 May 2020 | Published 1 September 2020



## Abstract

With the increase in the popularity of mobile devices, malicious applications targeting Android platform have greatly increased. Malware is coded so prudently that it has become very complicated to identify. The increase in the large amount of malware every day has made the manual approaches inadequate for detecting the malware. Nowadays, a new malware is characterized by sophisticated and complex obfuscation techniques. Thus, the static malware analysis alone is not enough for detecting it. However, dynamic malware analysis is appropriate to tackle evasion techniques but incapable to investigate all the execution paths and also it is very time consuming. So, for better detection and classification of Android malware, we propose a hybrid approach which integrates the features obtained after performing static and dynamic malware analysis. This approach tackles the problem of analyzing, detecting and classifying the Android malware in a more efficient manner. In this paper, we have used a robust set of features from static and dynamic malware analysis for creating two datasets i.e. binary and multiclass (family) classification datasets. These are made publically available on GitHub and Kaggle with the aim to help researchers and anti-malware tool creators for enhancing or developing new techniques and tools for detecting and classifying Android malware. Various machine learning algorithms are employed to detect and classify malware using the features extracted after performing static and dynamic malware analysis. The experimental outcomes indicate that hybrid approach enhances the accuracy of detection and classification of Android malware as compared to the case when static and dynamic features are considered alone.

## Keywords

Android Malware, Dynamic Malware Analysis, Machine Learning, Static Malware Analysis.

DOI: 10.9781/ijimai.2020.09.001

## I. Introduction

SMARTPHONES have become an open source platform for running different types of applications (apps) such as banking, lifestyles, gaming, education, etc. According to the site-worldwide mobile application, download of apps reached 205.4 billion in year 2018 and will increase continuously [1]. The fast growth in the smartphone industry has made lot of users to use smartphones to consume multiple services and access the Internet. The Android apps bring lot of comfort for our life by supporting persistent communication everywhere and also providing diverse functionalities. The expansion of Android apps plays a vital role for the progress of upcoming economy and mobile Internet.

The smartphones usually store user's private data such as messages, pictures and personal information etc. As a result, these smartphones become the target of attackers [2], [3]. Nowadays in smartphone industry, Android operating system (OS) has gained the highest position throughout the world. In 2018, the wide use of Android apps has resulted in an increase of Android malware (approximately 2.84 million) [4]. According to the report of McAfee, 31 million Android malware were found in 2018 and also shows that approximate 1.9 million new samples are identified every year [5]. As a result, it has become complicated to manually process large amount of Android malware samples. Thus, it becomes a most challenging task for antivirus companies to detect and classify malware. To evade the problem of handling large amount of malware samples manually and the malware obfuscation, the researchers start finding efficient techniques of Android malware detection and family classification.

The researchers are making use of several methods for detection of Android malicious apps. The traditional method to identify Android malware is relying on a signature based technique in which the signature of an app is matched with the already existing signatures present in the database. The major limitation of this technique is that it cannot identify unfamiliar malware. The ongoing research for detection and classification of malware is based on two methods i.e. static and dynamic malware analysis [2]. Static malware analysis method examines the code of the app to detect the malicious patterns without running the code [6]. It provides fast detection and high efficiency. But this method fails to identify the Android apps which make use of code obfuscation techniques [7]. The dynamic malware analysis method investigates the behavior of app while executing in a virtual environment. It is more efficient but this method is resource and time intensive. Moreover, this type of analysis is incapable to investigate all the execution paths. In order to strengthen the

* Corresponding author.
E-mail address: ekta.gandotra@gmail.com

accuracy, the features acquired from both static and dynamic analysis can be integrated [8]. Moreover, there exists only limited benchmark datasets available publically to evaluate the proposed machine learning techniques.

In this paper, we have worked on both detection and family classification of Android malware. Here detection relates to a binary classification problem which consists of two classes "malware" and "benign" and family classification relates to the multiclass classification problem which consists of 13 malicious families. Android malware family signifies a group of malicious programs that share common behavior and are generated from the same source code. We propose a hybrid approach for detection and classification of Android malicious apps. It depends on the fusion of static and dynamic malware analysis. Initially, we perform static malware analysis for extracting static features based on API calls, command strings, permissions and intents. Then, we performed dynamic malware analysis to extract features using CuckooDroid [9]. CuckooDroid is an extension of cuckoo sandbox which is used for automatic analysis of Android suspicious files [10]. The features considered for dynamic malware analysis are based on cryptographic operations, dynamic permissions, information leaks and system calls. In order to strengthen the accuracy, we integrate the features acquired from both static and dynamic malware analysis. Considering the presence of irrelevant, noisy and redundant features, an information gain ranking algorithm is applied to extract the relevant features.

### A. Research Contributions

The major contributions of the paper are as follows:

1. Two datasets i.e. binary and multiclass (family) classification datasets are created (using static and dynamic malware analysis) and shared publically on GitHub and Kaggle.
2. Feature selection method is used to choose the appropriate set of features for both the datasets.
3. The relevant features selected for both static and dynamic malware analysis are integrated.
4. Machine learning (ML) algorithms belonging to different categories are employed and evaluated on both the datasets for static, dynamic and integrated features.

### B. Organization

The rest of the paper is structured as follows: section II summarizes the related work on classification and identification of Android malware. Section III describes the proposed methodology. Section IV demonstrates the experimental outcomes based on different evaluation parameters. Section V concludes the paper and provides future scope.

## II. Related Work

In the literature, researchers have developed various novel techniques for identification and classification of Android malware using ML methods. Current malware identification methods fall under two categories i.e. static and dynamic malware analysis [11]. This section discusses the work associated with malware detection and classification based on static and dynamic malware analysis using ML methods.

### A. Static Malware Analysis

The static malware analysis is the way to discover the malicious patterns in app by examining its code. In order to find out the malicious patterns [12], it uses disassemble techniques to decompile the app source code [13]. This subsection includes the research papers related to static malware analysis which focuses on detection and classification of Android malware.

Li et al. [14] suggested a malware identification system known as significant Permission Identification (SigPID). They build 3 levels of pruning by extracting permission data to determine the relevant permissions that can be to distinguish between malware and benign apps. The authors employed ML methods to classify the Android apps. The experimental results show that SigPID performs better with 93.62% of accuracy as compared to existing approaches. In [15], the authors suggested a highly efficient method to extract API calls, permission-rate, surveillance system events and permissions as features. They constructed a model based on ensemble Rotation Forest to identify whether an app is malicious or benign. The results demonstrate that the proposed approach obtained highest precision of 88.16% with 88.26% accuracy at the sensitivity of 88.40%. Yerima and Sezer [16] introduced a novel fusion technique (DroidFusion) which includes amalgamation of various ML techniques for improving accuracy. The DroidFusion creates a model by training classifiers and then they employed a feature ranking algorithm on the predictive accuracies in order to acquire a final classifier. The results indicate that DroidFusion is more superior than stacking ensemble method. In [17], the authors presented a multimodal deep learning based framework for the identification of Android malware. They extracted diverse features and refined these using similarity based or existence-based method. The results show that the accuracy obtained by the multimodal deep learning framework is 98%. Feizollah et al. [18] presented an analysis of the usefulness of intents for classifying the malicious apps. They reported that intents are more important feature than permissions for classification of malware. The results demonstrate that detection rate of intent and permission is 91% and 83% respectively. The authors also indicate that the detection accuracy of combined features is 95.5% which is higher than the individual features. In [19], the authors explored the risk based on permissions in Android apps. They applied T-test, correlation coefficient and mutual information to rank the specific permission according to their risk. Principal component analysis and sequential forward selection are employed to determine the subsets of risky permission. They evaluated the effectiveness of risky permission for detection of malapp with Decision Tree (DT) Support Vector Machine (SVM) and Random Forest (RF). The results indicate that the detection accuracy of malapp detector is 94.62% with 0.6 False Positive Rate (FPR). Dhalaria et al. [20] performed a comparative analysis between different base classifiers such as SVM, Logistic Regression (LR), Naive Bayes (NB) K-Nearest Neighbor (K-NN), DT, RF and ensemble techniques (Bagging, Stacking and Boosting). The experimental results demonstrate that the stacking ensemble technique found to be more superior then the base classifiers. Dhalaria et al. [21] employed a convolutional neural network (CNN) to classify Android malicious apps. The grayscale images of classes.dex and AndroidManifest.xml are created which are extracted from the Android package. The experimental results indicate that the classes. dex file performs better in comparison to AndroidManifest.xml.

The static malware analysis is quicker in analyzing the code but it fails against code obfuscation techniques and morphed malware. The dynamic malware analysis overwhelms the constraints of static malware analysis.

### B. Dynamic Malware Analysis

It executes the samples in runtime environment such as an emulator and a virtual machine to track the behavior of the app. This section includes the literature on detection and classification of Android malware using dynamic malware analysis.

Cai et al. [22] presented a novel classification approach (DroidCat) which is based on dynamic analysis. The authors used a set of dynamic features such as method calls, app resources and Inter-Component Communication. The experimental outcomes indicate that DroidCat

obtained 97% accuracy and F-measure for classifying the Android malicious apps. In [23], the authors proposed a dynamic analysis framework i.e. EnDroid which used different types of dynamic features for the identification of malware. They employed a chi-square algorithm to select the relevant features and applied an ensemble learning technique to differentiate between malware and benign apps. Das et al. [24] proposed the model named as frequency centric for feature construction using system calls to effectively identify the malware. The authors build a ML method using Multilayer Perceptron (MLP) in FPGA in order to train a classifier. They found that the proposed approach obtained low power consumption, fast detection and high accuracy. In [25], the authors addressed TaintDroid, a dynamic taint tracking which is proficient of continuously tracking various source of sensitive data. As a result, it provides security service firms seeking and essential input for Android users to identify malicious apps. Chen et al. [26] presented a framework which uses a classification scheme named as Model-Based Semi-Supervised (MBSS). The authors also compared their proposed approach with the existing approach such as K-NN, Linear Discriminant Analysis (LDA) and SVM. The results indicate that the proposed approach achieves 98% accuracy at very low FPR. In [27], the authors designed and implemented a dynamic analysis method named as DroidTrace. It examined the system calls which are executed in dynamic payloads. DroidTrace also carried out physical alteration to trigger numerous dynamic loading behaviors within an app.

The dynamic malware analysis can detect the unfamiliar malware that a static analysis cannot but it takes more time and resources. Moreover, it explores only a single execution path.

### C. Hybrid Malware Analysis

Gandotra et al. [8] suggested that single approach either dynamic or static is not sufficient for accurately classifying the malware due to the obfuscation and execution stalling. To overcome this problem, the researchers have started to make use of a hybrid analysis approach. This section includes the work done in the field of hybrid malware analysis which focuses on detection and classification of Android malware.

Yuan et al. [28] introduced an engine named as DroidDetector which automatically characterized the app as either malware or benign. The authors extracted the features using static and dynamic analysis. The experimental results demonstrate that DroidDetector obtained highest accuracy 96.76% when compared with conventional ML techniques. In [29], the authors proposed the hybrid approach for identification of malware using static and dynamic analysis. They created the normal and malicious pattern sets by matching the pattern of benign and malware apps with each other. To determine the unknown app, the authors also compared these with both normal and malicious pattern sets offline. The results demonstrate that the proposed approach obtained better detection rate. Martin et al. [30] presented an OmniDroid dataset consisting of 22,000 malware and benign samples. They developed a framework for static and dynamic analysis of apps and applied ensemble learning classifiers for identification of malicious apps. In [31], the authors presented an Android Application Sandbox (AASandbox) which is capable to carry out both dynamic and static analysis to identify malicious apps. For providing distributed and fast detection, they deployed the detection algorithm and sandbox in the cloud. The results show that AASandbox is more efficient than antivirus apps available for Android OS.

From the literature survey, it is found that the hybrid approach is capable to classify the Android apps more accurately. Though, a lot of work has been reported in the literature on detection (binary classification) of Android apps using hybrid approach but the least focus has been paid on family classification of Android malware. Moreover, there exist only two benchmark datasets i.e. Malgenome [3] and Derbin [32] which have been made public over past few years. These datasets include old Android apps and were created in the years 2012 and 2014 respectively. But nowadays, evolving malwares are so sophisticated and complex that they cannot be recognized easily. This paper presents the approach used for creating our own datasets. These consist of recent Android apps and we have made these publically available on GitHub and Kaggle. These would help the research community to evaluate their proposed ML techniques for malware classification. Different machine learning algorithms are employed on these two datasets to perform binary and family classification of Android apps when both static and dynamic features are integrated.

## III. Proposed Methodology

This section discusses the proposed methodology for detection and family classification of Android apps. It consists of three phases i.e. data collection, data preparation and detection & family classification. In the first phase, data is collected from various sources such as virusshare [33], apkmirror [34] and apkpure [35]. In the second phase, MD5 hash is applied to remove the duplicate apps and then these apps are examined using Avira Antivirus (AV) tool [36]. The static and dynamic malware analysis is performed to extract features from the Android apps. Static features are extracted using self-developed python script which uses multiple automated tools such as Baksmali Diassembler [37], String [38] and AXMLPrinter2 [39]. The features extracted using static malware analysis includes API calls, command string, permissions and intents. Dynamic features are extracted using CuckooDroid [9] which analyzes the behavior of app during runtime. The features extracted using dynamic malware analysis include dynamic permissions, cryptographic operations, information leaks and system calls. After feature extraction, an information gain feature ranking algorithm is employed in order to remove the noisy, irrelevant and redundant features. Various ML classifiers such as SVM, DT, RF, NB, K-NN PART and MLP are employed to identify and classify the Android apps. Fig. 1 shows the workflow of the proposed methodology.

### A. Data Collection (Phase-I)

The initial phase of the proposed methodology is data collection. The Android apps are collected from multiple sources such as apkpure, apkmirror and virusshare. These apps are stored in Android application packages (.apk) file format. A total of 4400 recent Android apps are downloaded from these sources. The malicious apps are downloaded from virusshare after getting registered with their website and also getting permission from the administrator. The benign apps are collected from apkpure and apkmirror.

### B. Data Preparation (Phase-II)

This subsection discusses various steps used for data preparation. These include removing duplicate applications, labelling, feature extraction and feature selection.

#### 1. Removing Duplicate Applications

MD5 hash algorithm is employed on the collected Android apps to eliminate the duplicate ones. After removing the duplicates, we are left with 3547 Android apps.

#### 2. Labelling

The unique Android apps obtained from the previous step are scanned using Avira Antivirus (AV) tool for labelling. After labelling, out of 3547 apps, 1747 are malicious and 1800 are benign. Furthermore, 1747 malicious apps are further labelled as 13 malware families as shown in Fig. 2.

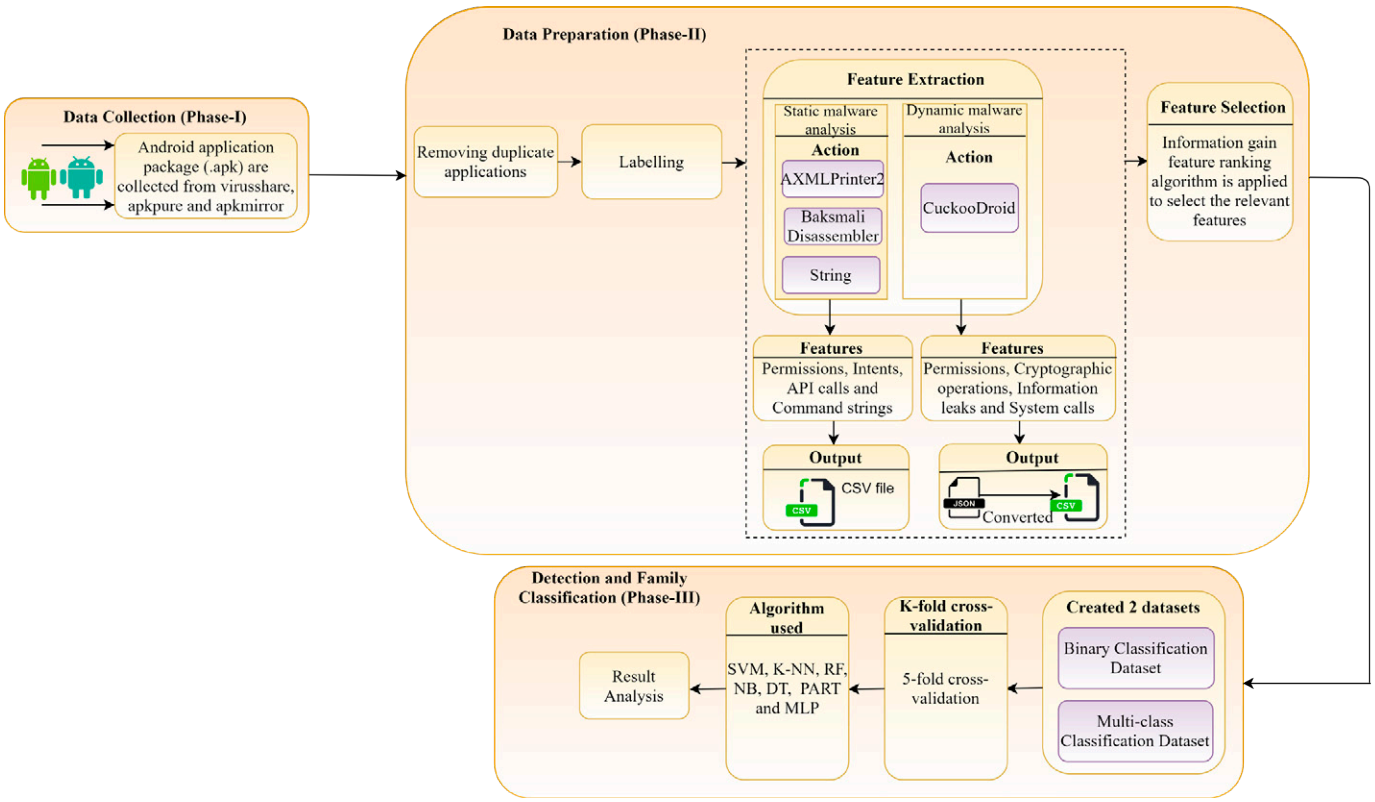



Fig. 1. Workflow of the proposed methodology.

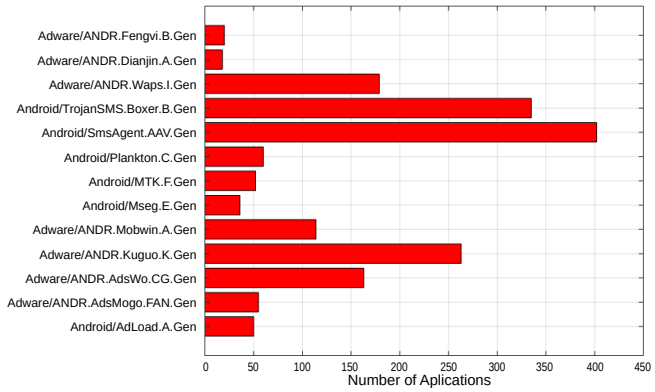



Fig. 2. Graphical representation of Android malware families.

## 3. Feature Extraction

Various features are extracted using static and dynamic malware analysis. In static malware analysis, we have extracted four different types of static features i.e. API calls, intents, permissions and command strings using self-developed python script which uses several automated tools such as Baksmali Disassembler, AXMLPrinter2 and string. In dynamic malware analysis, we have extracted four different types of dynamic features i.e. cryptographic operations, dynamic permissions, information leaks and system calls using CuckooDroid (Android malware analysis tool). The detailed description related to feature extraction using static and dynamic malware analysis is explained below.

### a) Using Static Malware Analysis

It is performed without executing the code. It uses various disassemble techniques to decompile the app source code. To extract the static features, we developed a python script which uses various automated tools i.e. Baksmali Disassembler, AXMLPrinter2 and string.

The features extracted for analysis using these tools are API calls, permissions, intents and command strings. The process of extracting features is shown in Fig. 3. The .apk file is saved in compressed zip format. To view the content of .apk file, we first need to unzip or unpack it. The .apk file consists of classes. dex file, Android Manifest file, res, lib and assets folder. Through this, we extracted four different types of static features using different static tools. Classes.dex file contains information about API calls, Android Manifest file contains information about permission and intents and the rest contains information about command strings. These features are selected on the basis of existing literature and the official site of Android which says that these specific features are more prominent in malicious applications [16], [40].

- **API calls**: It is used to interact with the device. These contain the method, classes and packages to help developers to build apps. The Android is based on java programming language and Java compiler converts the source code into java bytecode. It uses Dalvik Virtual Machine (DVM) after disassembling java bytecode, it gives information about packages, methods and classes. A total of 47 API calls are extracted using a self-developed python script after decompiling classes.dex with Baksmali Disassembler.
- **Permissions**: The main purpose of permissions is to secure the privacy of the users. The apps must request permission to access user sensitive information and system features. The system sometimes gives permission itself or could provoke users to accept the request. Permission is mainly declared in the AndroidManifest. xml. A total of 277 permissions are extracted using a self-developed python script after decompiling AndroidManifest.xml with AXMLPrinter2.
- **Command strings**: It is one of the static features which is used for identification of Android malware. It analyzes the command string which is present in lib, res, assets folder. A total of 6 command strings are extracted using a self-developed python script after decompiling lib, res and assets with string.

- **Intents**: Intents are found in Manifest.xml. It infers the intentions of apps e.g. pick a contact, dial a number etc. Intents are extracted from manifest.xml after decompiling with AXMLPrinter2. A total of 22 intents are extracted using a self-developed python script after decompiling AndroidManifest.xml with AXMLPrinter2. Table I lists some of the examples of static features considered under these four categories.

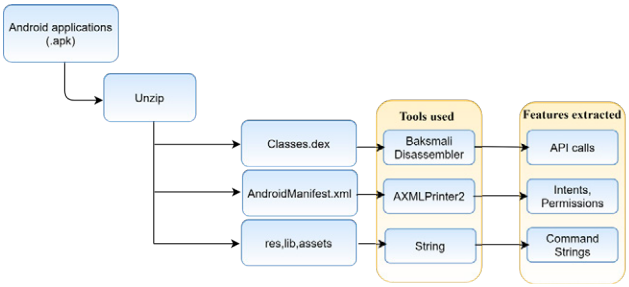



Fig. 3. Process of extracting static features.

### *b) Using Dynamic Malware Analysis*

It is performed while executing the code in the runtime environment. The runtime behavior information of the apps is obtained using the open source dynamic analysis tool named as CuckooDroid. It is an extension of cuckoo sandbox, the open source software for executing and analyzing the apps. It automatically executes and analyzes files and collects the information of the file at runtime. CuckooDroid is liable for handling the Android emulator and produce report at the termination of analysis. Cuckoo's infrastructure consists of a guest machine (i.e. the virtual machine that carry out analysis) and the host machine (i.e. the management software). The host runs the main components of the sandbox that controls the whole analysis process, whereas the guest machine is the isolated environment where the Android malware samples are carried out. The guest machine consists of Linux virtual machines that run Android emulator, which is monitored by the machinery module. The main work of Android emulator is to carry out the execution of apps, collect information and report it back to CuckooDroid. Every Android malicious file is run until all processes are finished or a timeout of 180 seconds is reached which means an Android sample is given a maximum of 180 seconds for analysis. After the analysis of particular sample is over, the results are compiled in JSON format. We need a guest machine which is to be rooted Android Virtual Device (AVD) with xposed framework [41] and with its two module i.e. Emulator Anti-Detection and Droidmon. Python 2.7 is used to run the analyzer code and python agent on guest machine. The role of the python agent is for analysing code, receiving APK file, and carrying out the analysis. The python analyzer executes apps, send screenshots back to host, send dropped files back to host. It is liable for terminating the analysis and sending back some log file to host. After the complete procedure, the log reports are collected which is in the Java Script Object Notation (JSON) format. The reports produced by Cuckoo Droid for different apps are then parsed and saved to the database in CSV format using Python script. Afterwards, these are used for detection and classification of malware. The process of extracting dynamic features is shown in Fig. 4. The features extracted for analysis are cryptographic operations, information leaks, dynamic permissions and system calls.

These features are selected on the basis of existing literature and the official site of Android which says that these specific features are more prominent in malicious applications [22], [40], [42]. The detailed description of these four features is explained as follows:

- **Cryptographic operations**: Malware accepts these operations to target premium sms number, encrypt root exploits, malicious payload etc. To distinguish various cryptographic behaviors, these features are formed as <action>_<algorithm >. Here <action> includes various operations like key generation, decryption and encryption and the <algorithm> includes various cryptographic algorithms. A total of 79 cryptographic operations are extracted using CuckooDroid.

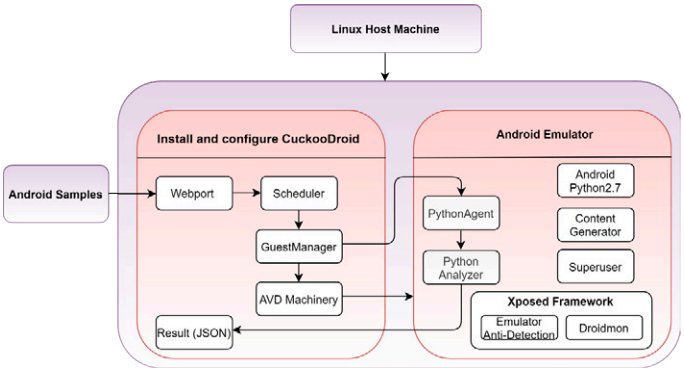



Fig. 4. Process of extracting dynamic features.

- **Dynamic permissions**: It is considered as one of the important dynamic features to analyze the behavior of apps. Dynamic permissions are those permissions which are executed at the runtime environment. A total of 71 dynamic permissions are extracted at runtime using CuckooDroid.
- **Information leaks**: Confidential and personal data has newly gained more attention. Malware usually vigorously harvests numerous data on contagious devices, such as contact information, IMEI, SMS contents, credential information related to social network and banking etc. The collected data may be used to make profits, keep track on users and acquire authorized account etc. These features are defined as <source>_<sink>. Here <source> includes operations gaining confidential data and the <sink> includes operations leaking confidential data. A total of 123 information leaks are extracted at runtime using CuckooDroid.
- **System Calls**: It is one of the most important dynamic features of Android app. It is an efficient feature for intrusion detection in a mobile device. Through system calls, Android apps take services of the kernel. The kernel offers useful functions to apps such as device security, process related to operations and power management etc. These malware usually invokes sigprocmask, getuid, ptrance to affect the execution of other apps. A total of 50 system calls are extracted at runtime using CuckooDroid. Table II lists some of the examples of dynamic features considered under these four categories.

After performing static and dynamic malware analysis, a total of 352 static and 323 dynamic features are extracted from all the Android apps considered in this work. Thus, we have come up with two datasets. First is a binary classification dataset consisting of 1747 malicious and 1800 benign apps. Second is a multiclass classification dataset consisting of 1747 malicious apps belonging to 13 malware families. Both these datasets are made public on GitHub and Kaggle (Link: https://github.com/Meghna-Dhalaria/Android-malware-dataset) and (Link: https://www.kaggle.com/meghnadhalaria/android-malware-detection-and-classification) respectively.

## *4. Feature Selection*

It is also known as attribute selection. It is used for dimensionality reduction which helps in choosing relevant features. Irrelevant and redundant features can decrease the quality of the classification model and the accuracy. Higher dimensional datasets required more space and computation time [43]. Selecting the relevant features will help

TABLE I. Examples of Static Features Considered

| Features | Number of features | Examples | Feature value |
|---|---|---|---|
| **API Calls** | 47 | onserviceConnected, Ljavax.crypto.spec.SecretKeySpec, getBinder, android.os.Binder, Ljava.net.URLDecoder, ServiceConnection, KeySpec, Ljava.lang.Class.getMethods | If an API call (out of 47) is existing in the classes.dex then the value of that feature is set to 1 otherwise 0. |
| **Permissions** | 277 | GET_TASKS, READ_PHONE_STATE, WRITE_EXTERNAL_STORAGE, RECEIVE_BOOT_COMPLETE, READ_SMS, SYSTEM_ALERT_ WINDOW, RECEIVE_SMS, ACCESS_NETWORK_STATE | If a permission (out of 277) is existing in the Manifest.xml file then the value of that feature is set to 1 otherwise 0. |
| **Command Strings** | 6 | Chown, /system/bin, mount, /system/app, remount | If a command string (out of 6) is existing in the *res, lib, assets* folder then the value of that feature is set to 1 otherwise 0. |
| **Intents** | 22 | CALL_BUTTON, SET_WALLPAPER, NEW_OUTGOING_CALL, SCREEN_OFF, PACKAGE_CHANGED, ACTION_SHUTDOWN, BATTERY_LOW | If an intent (out of 22) is existing in the Manifest.xml file then the value of that feature is set to 1 otherwise 0. |

TABLE II. Examples of Dynamic Features Considered

| Features | Number of features | Examples | Feature value |
|---|---|---|---|
| **Cryptographic Operations** | 79 | Decryption_AES, encryption_AES, keyalgo_AES | If a cryptographic operation (out of 79) is present in JSON file then the value of that feature is set to 1 otherwise 0. |
| **Dynamic Permissions** | 71 | AUDIO_FILE_ACCESS, ACCESS_ GOOGLE_ PASSWORDS, WRITE_ CONTACT_DATA, READ_CONTACT_DATA | If a dynamic permission (out of 71) is present in JSON file then the value of that feature is set to 1 otherwise 0. |
| **Information Leaks** | 123 | IMEI_File, IMSI_Network, IMSI_File, PHONE_NUMBER_File, IMEI_ Network | If an information leak (out of 123) is present in JSON file then the value of that feature is set to 1 otherwise 0. |
| **System Calls** | 50 | ptrace, recvfrom, sigprocmask, write, wait4, sendto, getpid, read, recvmsg, chmod, sendmsg | If a system call (out of 50) is present in JSON file then the value of that feature is set to 1 otherwise 0. |

in reducing the space and time complexity and also help in increasing the accuracy. In this work, we have employed an information gain feature ranking algorithm [44] to select the relevant features for better detection and classification of Android malware. Information gain calculates the quantity of information provided about the class. It makes use of entropy to compute the homogeneity of samples. The entropy $H(X)$ of the dataset (having c number of classes) is calculated as given in equation (1).

$$H(X) = \sum_{i=1}^{c} -p_i \log_2 p_i \quad (1)$$

Where $p_i$ is the probability of class $i$ in the dataset $X$. The dataset is then split on the different attributes $A$. The entropy for a dataset with respect to attribute $A$ i.e. $H(X, A)$ is calculated using equation (2).

$$H(X, A) = \sum_{k \in A} P(c) H(c) \quad (2)$$

Here $k$ represents the possible values of the attribute $A$.

Information gain achieved by an attribute is expressed as shown in equation (3). Greater the Information Gain (IG) of a particular feature, more important the feature is.

$$IG = H(X) - H(X, A) \quad (3)$$

The information gain method assigns rank and weight to each feature. We have not considered the attributes with zero weight. Thus out of 352 features, we are left with 110 static features for binary classification dataset (named as Dataset-1) and 47 static features for family classification dataset (named as Dataset-2). Fig. 5 and Fig. 6 show the top 20 selected attributes for detection (Dataset-1) and family classification (Dataset-2) datasets respectively.

The datasets created using dynamic malware analysis consist of 323 features. Out of 323 features, we are left with 99 dynamic features in Dataset-1 and 35 features in Dataset-2. Fig. 7 and Fig. 8 show the top 20 selected dynamic features for detection (Dataset-1) and family classification (Dataset-2) datasets respectively.

The summary of both the datasets i.e. Dataset-1 and Dataset-2 before and after feature selection is given in table III. Fig. 9 shows the various steps for preparing these two datasets.

TABLE III. Description of Dataset (Where, # Stands for Number of)

| Dataset Name | #Benign apps | #Malicious apps | #Feature extracted | | #Feature selected | |
|---|---|---|---|---|---|---|
| | | | Static | Dynamic | Static | Dynamic |
| **Dataset-1** | 1800 | 1747 | 352 | 323 | 110 | 99 |
| **Dataset-2** | ----- | 1747 (with 13 families) | 352 | 323 | 47 | 35 |

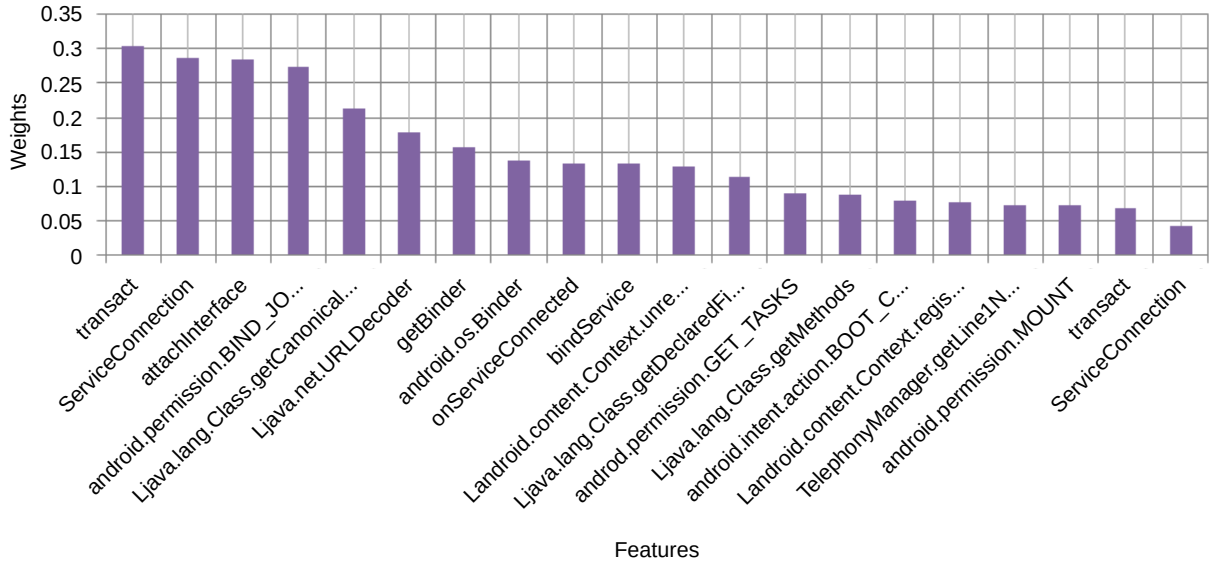



Fig. 5. Top 20 selected static features for detection dataset (Dataset-1).

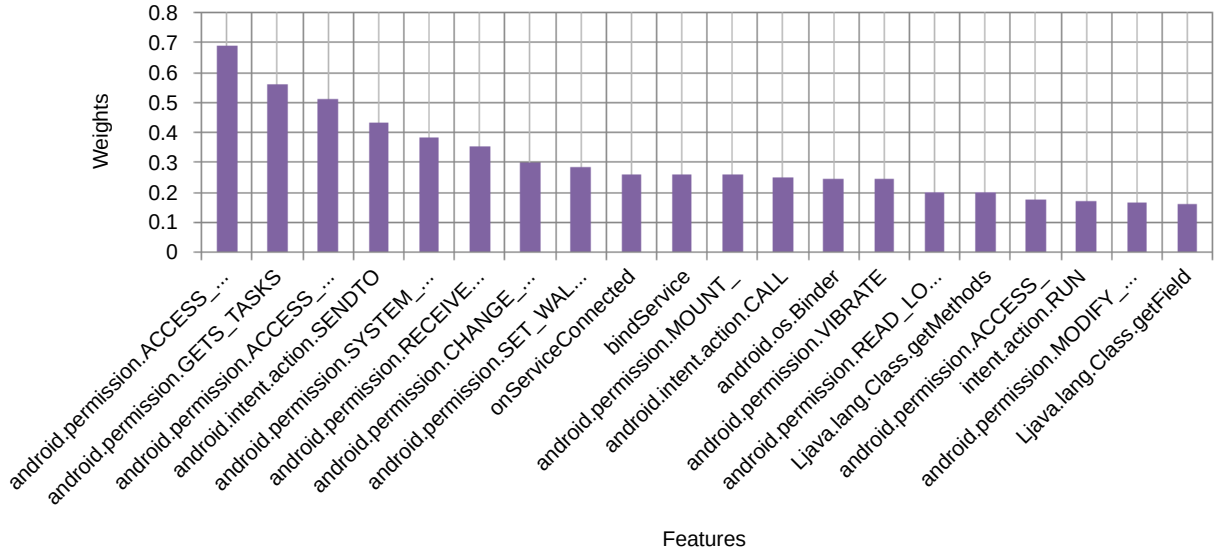



Fig. 6. Top 20 selected static features for family classification dataset (Dataset-2).

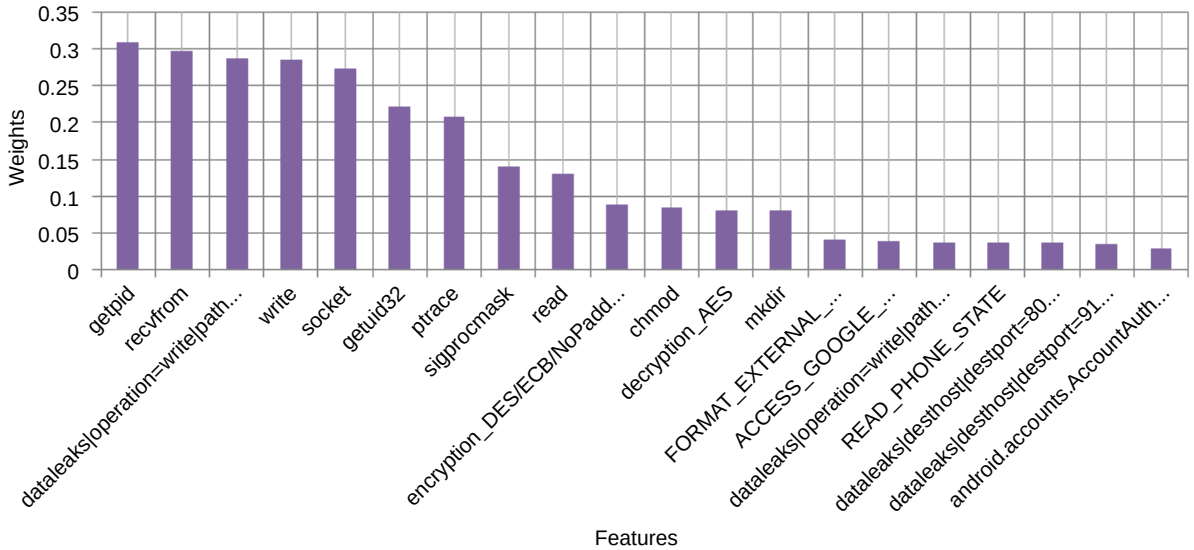



Fig. 7. Top 20 selected dynamic features for detection dataset (Dataset-1).

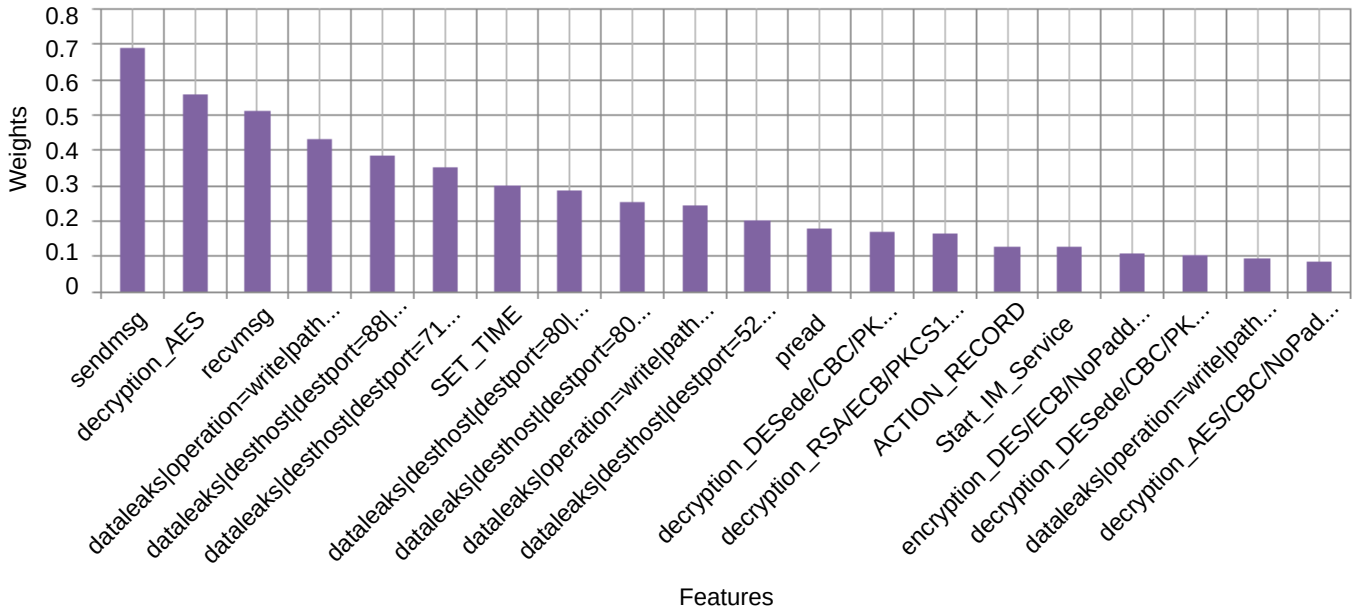



Fig. 8. Top 20 selected dynamic features for family classification dataset (Dataset-2).

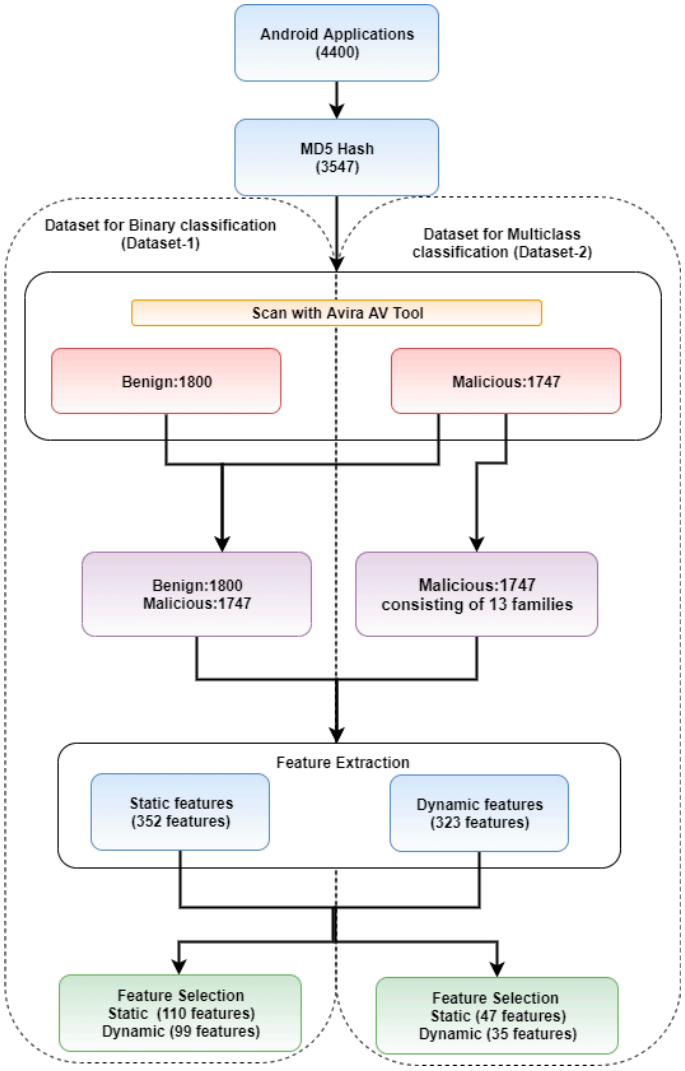



Fig. 9. Process of Data Preparation.

### C. Detection and Family Classification (Phase-III)

Various ML algorithms i.e. SVM, RF, DT, NB, K-NN, PART and MLP are used to build models for detection and classification of Android malware. These models are trained using 5-fold cross validation, in which the whole dataset is divided into 5 equal parts. Four parts are used to train the model and the remaining part is used for testing at every run. This section provides the brief introduction of ML algorithms and the evaluation parameters used for evaluating these algorithms.

#### 1. Machine Learning Algorithms

The various ML algorithms used in this work are as follows:

- K-NN is one of the easiest supervised learning methods. It is also called as lazy learner [45]. This method does not depend upon the structure of data, whenever the new instance arises; it finds the closest training samples to the new instance by using distance measures such as Euclidean distance, Manhattan distance. At the end, by using the majority voting concepts it finds the class of the new instance.
- SVM is a method [46] which divides the data using a hyperplane. It acts like a decision boundary. It randomly draws the hyperplane and then computes the distance between the hyperplane and the closest data points (also called as support vector). It attempts to identify the optimal hyperplane that maximizes the margin.
- RF is an ensemble learning technique which involves a large number of individual decision trees that act as an ensemble [47]. Every decision tree produces a classification for input data and then RF collects the classification and illustrates the result based on majority voting.
- The structure of DT is like a tree, where non-leaf or internal node demonstrates a test on an attribute, topmost node represents the root node, terminal or leaf node holds a class label and the branch of the tree demonstrates the results of the test. In this work, we have used C4.5 algorithm to classify Android malware [48].
- The concept of NB is based on Bayes theorem. It forecasts the class membership probabilities i.e. the probability that a given tuples relates to an individual class. It is used for both binary and multiclass classification problems [49].

- PART is a partial decision tree algorithm. It is a separate and conquer rule learner. This technique produces sets of rules known as decision list. A new sample is compared to each rule and then the sample is assigned the class of the first matching rule [50].
- Multilayer Perceptron (MLP) is also called as Multilayer Neural Networks [51]. It consists of an input layer, an output layer and the hidden layer. It has various output units. The units of the hidden layer become input for the next layer. Semwal et al. [52], [53] worked in the field of different classification problems using deep learning techniques such as DNN based classifier and ANN. In [54], the authors [54] worked in the Extreme Machine Learning (ELM) for classification and prediction of gait data. In our work, we applied MLP for detection and classification of Android malware. We run the MLP for hidden layer h=3 and h=5 for Dataset-1 and Dataset-2 respectively. The activation function used for Dataset-1 and Dataset-2 are sigmoid and Softmax respectively. The learning rate is considered to be as 0.3. Fig. 10 shows the general framework of backpropagation based on neural network [53].

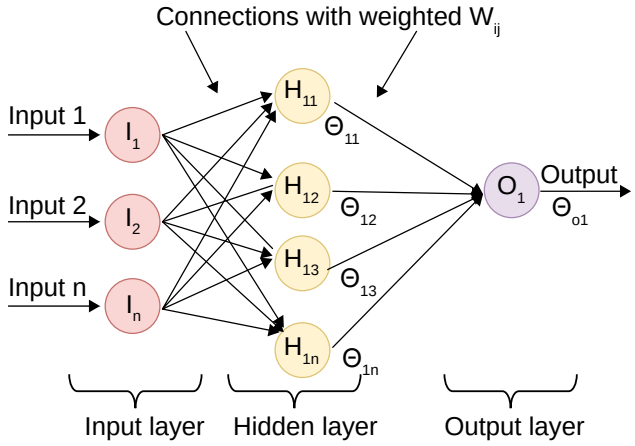



Fig. 10. General framework of backpropagation based on neural network [53].

The algorithm first initializes the weights to all nodes and then calculates the net input and output. It calculates the error rate and propagates it back. At the end, it updates the bias and weights and run the loop until the error becomes below the threshold.

### 2. Evaluation Parameters

The performances of the classifiers are assessed on the basis of various metrics such as precision, true positive rate (TPR), F-measure, false positive rate (FPR), Matthews correlation coefficient (MCC) and Area under curve (AUC) [55]. These performance metrics are defined using true negative (TN), false positive (FP), false negative (FN) and true positive (TP).

- **TPR**: It is also known as recall or sensitivity. It is defined as the ratio of true positive cases divided by the total number of actual positive cases. It is computed as shown in equation (4).

$$TPR = \frac{TP}{TP+FN} \quad (4)$$

- **FPR**: It is the ratio of false positive cases divided by total number of actual negative cases. It is computed as given in equation (5).

$$FPR = \frac{FP}{TN+FP} \quad (5)$$

- **Precision**: It is defined as the ratio of actual true predictive instances divided by the total number of true cases. It is computed as shown in equation (6).

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

- **F-measure**: It signifies the harmonic mean of recall and precision. It is calculated as shown in equation (7).

$$F-measure = \frac{2\times(Precision\times Recall)}{(Precision+Recall)} \quad (7)$$

- **Accuracy**: It is the ratio of true positive and true negative instances divided by the total number of instances. It is calculated as shown in equation (8).

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (8)$$

- **MCC**: It is used to measure the quality of binary classification algorithms. Its value lies between -1 to +1. Here -1 means inverse prediction and +1 means a perfect prediction. It is calculated as shown in equation (9).

$$MCC = \frac{TP\times TN-FP\times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}} \quad (9)$$

- **AUC curve**: It is one of the most significant parameters to measure the performance of classification models. It represents the measure of the separability.

## IV. Experimental Results

This section describes the experimental results based on static, dynamic and the hybrid features. Seven different ML technique are used which are run on python 3.7 under Intel Core i5 processor, 64 bit with 8GB RAM. We conducted the experiments using 5-fold cross validation method and evaluated the ML techniques on the basis of various evaluation parameters like TPR, F-measure, Accuracy, FPR, Precision, AUC and MCC.

### A. Classification Results Based on Static Features

Seven ML algorithms are used to detect and classify malware on detection (Dataset-1) and family classification (Dataset-2). These algorithms are carried out in python script through *sklearn* [56] library.

Table IV demonstrates the evaluation results of ML techniques on static malware analysis for Dataset-1. It shows that RF gives the best accuracy of 96.50% followed by K-NN and MLP with accuracy as 95.74% and 95.71% respectively.

Fig. 11 shows the comparison of different classifiers based on accuracy and MCC of static features for Dataset-1. It indicates that RF performs better in comparison to other classifiers. The accuracy and MCC obtained by RF is 96.50% and 0.933 respectively.

Table V shows the evaluation results of ML techniques using static features for family classification on Dataset-2. It is found that RF algorithm gives better accuracy i.e. 86.72% followed by SVM and DT which gives and accuracy of 85.86% and 84.77% respectively. The TPR, precision and F-measure obtained by RF is 0.867, 0.870 and 0.866 respectively which are better results than those obtained by other classifiers.

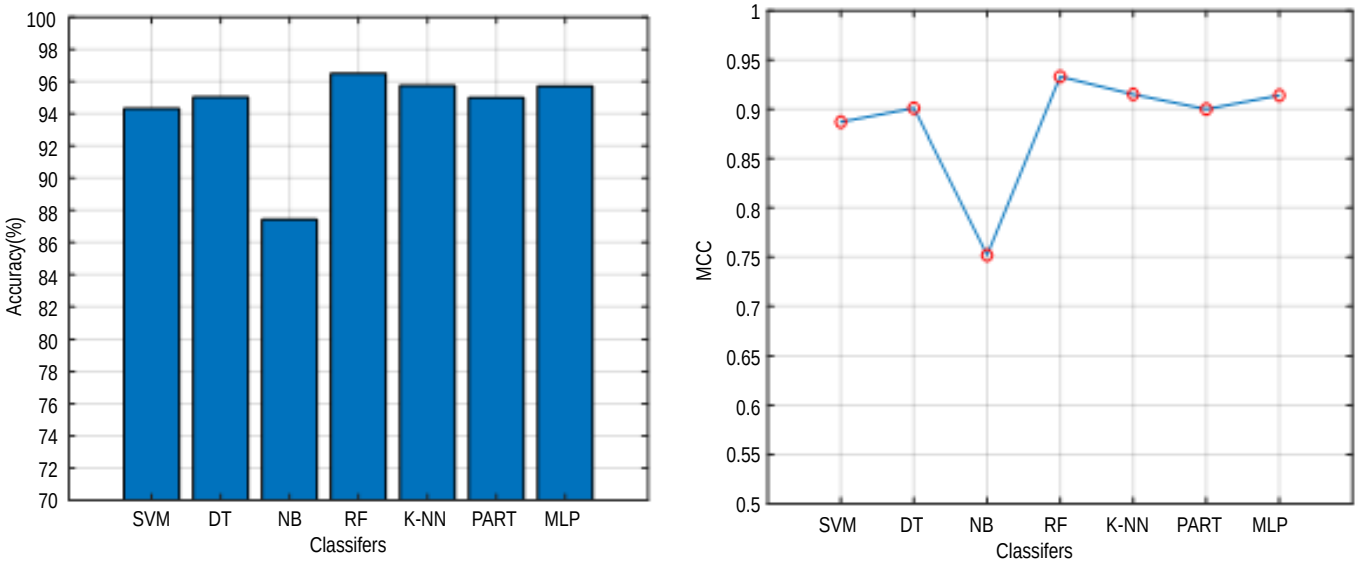
Fig. 12 shows the comparative analysis of different classifiers based on accuracy for Dataset-2. The maximum accuracy of 86.72% is obtained by RF. This value is much smaller than the results obtained in static malware analysis for detection of malware in case of binary classification.

TABLE IV. Classification Results Using Static Features for Dataset-1

| Classifiers | TPR | FPR | Precision | F-measure | MCC | AUC | Accuracy (%) |
|---|---|---|---|---|---|---|---|
| **SVM** | 0.943 | 0.057 | 0.943 | 0.943 | 0.887 | 0.943 | 94.33 |
| **DT** | 0.950 | 0.050 | 0.950 | 0.950 | 0.901 | 0.970 | 95.03 |
| **NB** | 0.874 | 0.124 | 0.878 | 0.874 | 0.752 | 0.948 | 87.42 |
| **RF** | 0.965 | 0.035 | 0.965 | 0.965 | 0.933 | 0.990 | **96.50** |
| **K-NN** | 0.957 | 0.042 | 0.958 | 0.957 | 0.915 | 0.989 | 95.74 |
| **PART** | 0.950 | 0.050 | 0.950 | 0.950 | 0.900 | 0.975 | 94.98 |
| **MLP** | 0.957 | 0.043 | 0.957 | 0.957 | 0.914 | 0.986 | 95.71 |

TABLE V. Classification Results Using Static Features for Dataset-2

| Classifier | TPR | FPR | Precision | F-measure | AUC | Accuracy (%) |
|---|---|---|---|---|---|---|
| **SVM** | 0.859 | 0.023 | 0.863 | 0.857 | 0.962 | 85.86 |
| **DT** | 0.848 | 0.023 | 0.852 | 0.847 | 0.949 | 84.77 |
| **NB** | 0.751 | 0.032 | 0.792 | 0.756 | 0.967 | 75.10 |
| **RF** | 0.867 | 0.024 | 0.870 | 0.866 | 0.982 | **86.72** |
| **K-NN** | 0.845 | 0.024 | 0.847 | 0.843 | 0.966 | 84.48 |
| **PART** | 0.840 | 0.024 | 0.842 | 0.839 | 0.947 | 84.02 |
| **MLP** | 0.830 | 0.026 | 0.832 | 0.830 | 0.964 | 82.99 |




(a) (b)

Fig. 11. Comparison of different classifiers based on (a) Accuracy (b) MCC using static features for Dataset-1.

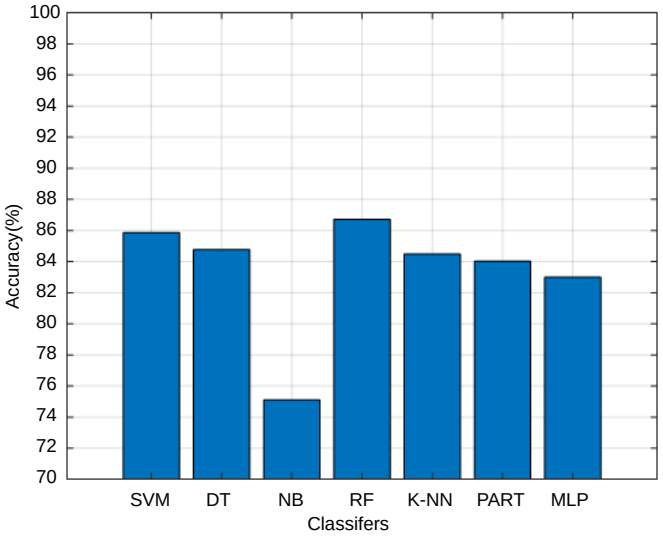



Fig. 12. Comparison of different classifiers based on accuracy using static features for Dataset-2.

### B. Classification Results Based on Dynamic Features

The static malware analysis is quicker in analyzing the code but it fails against code obfuscation techniques and morphed malware. So to overcome this problem, we considered the dynamic features for better detection and classification of malware. Seven ML algorithms are used to detect and classify malware on detection (Dataset-1) and family classification (Dataset-2) datasets.

Table VI shows the evaluation results of ML techniques on dynamic malware analysis for malware detection (binary classification) on Dataset-1. Among all these classifiers, RF is found to be more superior and accurate than other classifiers. The accuracy acquired by RF is 97.01% followed by SVM and MLP with 96.53% and 96.53% respectively.

Fig. 13 shows the comparative analysis of different classifiers based on accuracy and MCC using dynamic features for Dataset-1. It indicates that RF performs better in comparison to other classifiers. The accuracy and MCC obtained by RF is 97.01% and 0.940 respectively.

Table VII shows the evaluation results of ML techniques on dynamic malware analysis for family classification on Dataset-2. Among all these classifiers, RF is found to be more superior and accurate than other classifiers. The accuracy obtained by RF is 88.60% followed by SVM and DT with 86.85% and 84.25% respectively. The TPR, precision and F-measure obtained by RF is 0.886, 0.888 and 0.885 respectively which are better values than those obtained by other classifiers.

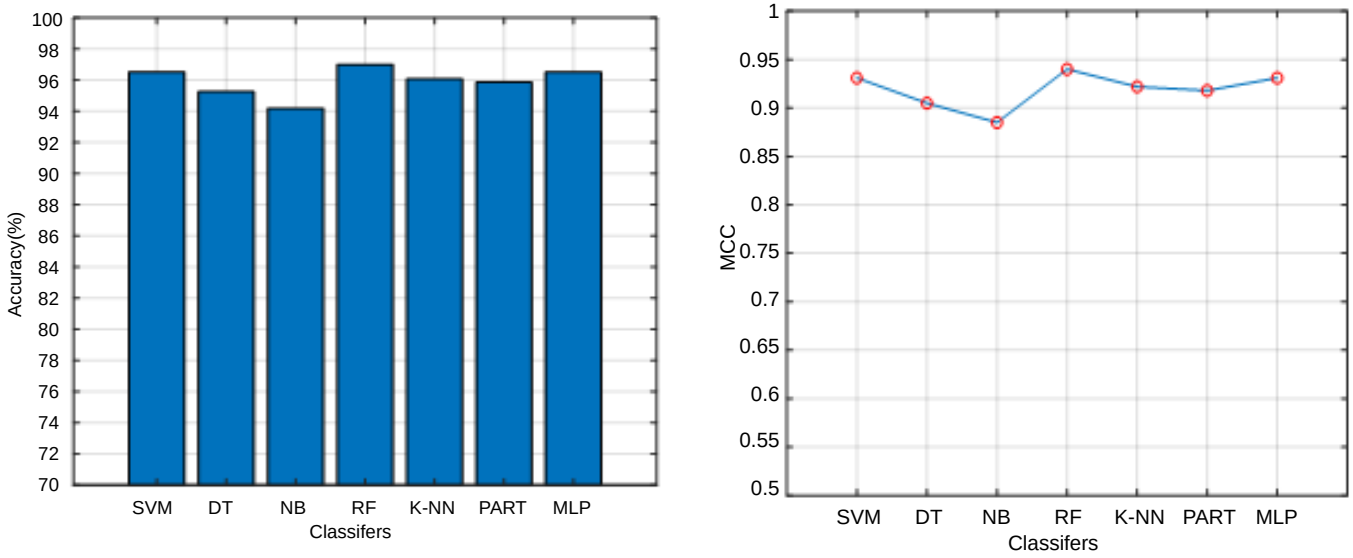
Fig. 14 shows the comparative analysis of different classifiers based on accuracy using dynamic features for Dataset-2. The maximum accuracy of 88.60% is obtained by RF. This value is much smaller than the results obtained in dynamic malware analysis for detection of malware (binary classification).

TABLE VI. Classification Results Using Dynamic Features for Dataset-1

| Classifier | TPR | FPR | Precision | F-measure | MCC | AUC | Accuracy (%) |
|---|---|---|---|---|---|---|---|
| **SVM** | 0.965 | 0.035 | 0.965 | 0.965 | 0.931 | 0.965 | 96.53 |
| **DT** | 0.953 | 0.048 | 0.953 | 0.953 | 0.905 | 0.973 | 95.26 |
| **NB** | 0.942 | 0.057 | 0.943 | 0.942 | 0.885 | 0.989 | 94.19 |
| **RF** | 0.970 | 0.030 | 0.970 | 0.970 | 0.940 | 0.996 | **97.01** |
| **K-NN** | 0.961 | 0.039 | 0.961 | 0.961 | 0.922 | 0.990 | 96.08 |
| **PART** | 0.959 | 0.041 | 0.959 | 0.959 | 0.918 | 0.970 | 95.88 |
| **MLP** | 0.965 | 0.035 | 0.965 | 0.965 | 0.931 | 0.988 | 96.53 |

TABLE VII. Classification Results Using Dynamic Features for Dataset-2

| Classifier | TPR | FPR | Precision | F-measure | AUC | Accuracy (%) |
|---|---|---|---|---|---|---|
| **SVM** | 0.864 | 0.021 | 0.871 | 0.866 | 0.985 | 86.85 |
| **DT** | 0.843 | 0.026 | 0.843 | 0.841 | 0.947 | 84.25 |
| **NB** | 0.800 | 0.029 | 0.805 | 0.795 | 0.951 | 79.96 |
| **RF** | 0.886 | 0.018 | 0.888 | 0.885 | 0.991 | **88.60** |
| **K-NN** | 0.839 | 0.025 | 0.842 | 0.837 | 0.967 | 83.91 |
| **PART** | 0.841 | 0.026 | 0.838 | 0.836 | 0.950 | 84.08 |
| **MLP** | 0.829 | 0.027 | 0.828 | 0.825 | 0.947 | 82.88 |




(a) (b)

Fig. 13. Comparison of different classifiers based on (a) Accuracy (b) MCC using dynamic features for Dataset-1.

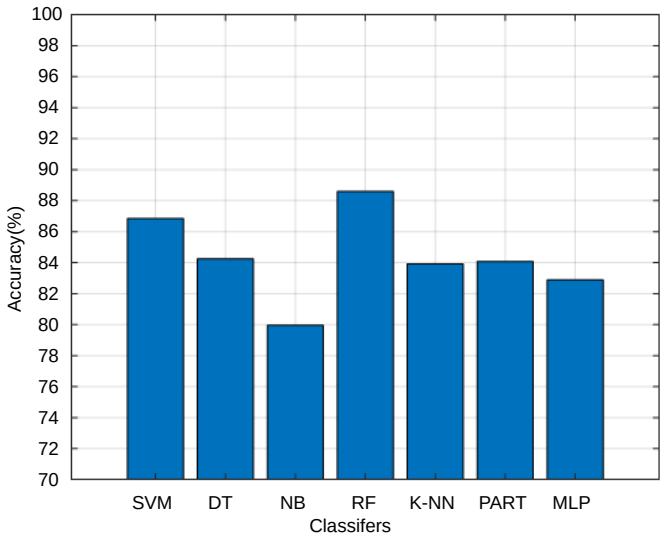



Fig. 14. Comparison of different classifiers based on accuracy using dynamic features for Dataset-2.

### C. Classification Results Based on Integrated Features

Single approach either static or dynamic is inadequate for correctly classifying the malware due to the obfuscation and execution stalling. So to overcome this problem, we make use of a hybrid analysis approach. We integrated the features obtained from both static and dynamic malware analysis. Seven ML algorithms are used to detect and classify malware on detection (Dataset-1) and family classification (Dataset-2) datasets.

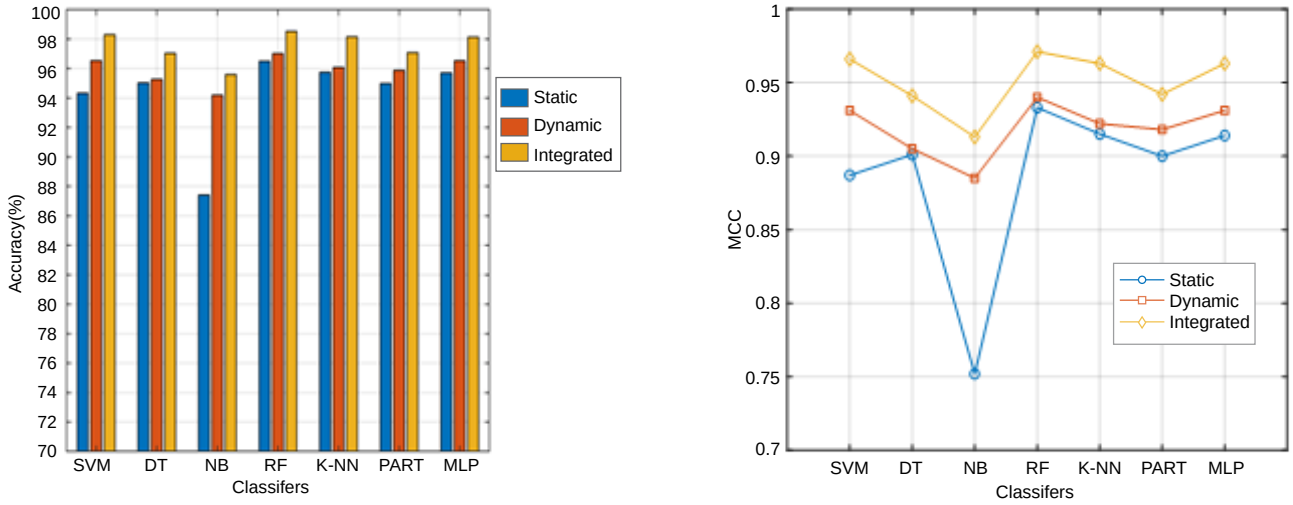
Table VIII shows the evaluation results of ML techniques on integrated features for Dataset-1. Among all these classifiers, RF is found to be more superior and accurate than other classifiers. The accuracy acquired by RF is 98.53% followed by SVM and K-NN with 98.30% and 98.16% respectively.

Table IX shows the evaluation results of ML techniques on integrated features for family classification for Dataset-2. Among all these classifiers, RF is found to be more superior and accurate than other classifiers. The accuracy acquired by RF is 90.10% followed by SVM and K-NN with 87.06% and 85.40% respectively. The TPR, precision and F-measure obtained by RF is 0.901, 0.902 and 0.901 respectively which are better results than those of other classifiers.

Fig. 15 shows the accuracy and MCC comparison of seven classifiers with respect to various approaches considered in our experiment for Dataset-1. It is clear from table VIII that there is an improvement in the accuracy and MCC for all the classifiers when the static and dynamic features are integrated. It means that using both static and dynamic features together helps for better detection and classification of the Android malware.

TABLE VIII. Classification Results Using Integrated Features for Dataset-1

| Classifier | TPR | FPR | Precision | F-measure | MCC | AUC | Accuracy (%) |
|---|---|---|---|---|---|---|---|
| **SVM** | 0.983 | 0.017 | 0.983 | 0.983 | 0.966 | 0.983 | 98.30 |
| **DT** | 0.970 | 0.030 | 0.970 | 0.970 | 0.941 | 0.980 | 97.03 |
| **NB** | 0.956 | 0.043 | 0.957 | 0.956 | 0.913 | 0.993 | 95.60 |
| **RF** | 0.985 | 0.015 | 0.985 | 0.985 | 0.971 | 0.999 | **98.53** |
| **K-NN** | 0.982 | 0.018 | 0.982 | 0.982 | 0.963 | 0.994 | 98.16 |
| **PART** | 0.971 | 0.029 | 0.971 | 0.971 | 0.942 | 0.983 | 97.09 |
| **MLP** | 0.981 | 0.019 | 0.981 | 0.981 | 0.963 | 0.993 | 98.13 |




(a) (b)

Fig. 15. Comparison of different classifiers based on (a) Accuracy (b) MCC using static, dynamic and integrated features for Dataset-1.

TABLE IX. Classification Results Using Integrated Features for Dataset-2

| Classifier | TPR | FPR | Precision | F-measure | AUC | Accuracy (%) |
|---|---|---|---|---|---|---|
| **SVM** | 0.870 | 0.020 | 0.875 | 0.871 | 0.987 | 87.06 |
| **DT** | 0.846 | 0.024 | 0.851 | 0.845 | 0.949 | 84.60 |
| **NB** | 0.783 | 0.027 | 0.814 | 0.784 | 0.970 | 78.30 |
| **RF** | 0.901 | 0.016 | 0.902 | 0.901 | 0.995 | **90.10** |
| **K-NN** | 0.854 | 0.022 | 0.857 | 0.854 | 0.966 | 85.40 |
| **PART** | 0.833 | 0.024 | 0.837 | 0.833 | 0.946 | 83.34 |
| **MLP** | 0.845 | 0.024 | 0.847 | 0.845 | 0.963 | 84.48 |

TABLE X. Classification Results of Best Classifier Using Static, Dynamic and Integrated Features for Dataset-1 and Dataset-2

| Dataset | Classifier | Approach | TPR | FPR | Precision | F-measure | MCC | Accuracy (%) |
|---|---|---|---|---|---|---|---|---|
| | | **Static** | 0.965 | 0.035 | 0.965 | 0.965 | 0.933 | 96.50 |
| **Dataset-1** | **RF** | **Dynamic** | 0.970 | 0.030 | 0.970 | 0.970 | 0.940 | 97.01 |
| | | **Integrated** | 0.985 | 0.015 | 0.985 | 0.985 | 0.971 | **98.53** |
| | | **Static** | 0.867 | 0.024 | 0.870 | 0.866 | -- | 86.72 |
| **Dataset-2** | **RF** | **Dynamic** | 0.886 | 0.018 | 0.888 | 0.885 | -- | 88.60 |
| | | **Integrated** | 0.901 | 0.016 | 0.902 | 0.901 | -- | **90.10** |

* MCC -- not applicable for multiclass dataset i.e. Dataset-2.

Fig. 16 demonstrates the comparison of seven classifiers on the basis of accuracy with respect to various approaches considered in our experiments for Dataset-2. It shows that for all the classifiers except NB and PART, the integrated approach performs better as compared to the cases when the static and dynamic features are considered alone. We are not able to achieve a good accuracy for the malware classification dataset (Dataset-2). It might be due to the imbalanced number of apps in different families.

Table X shows the comparison of static, dynamic and integrated approach for the best classifier i.e. RF for both the datasets i.e. Dataset-1 and Dataset-2. The results indicate that the integrated approach is found to be more appropriate for detection and classification of malware for both the datasets. The accuracy achieved by RF in case of Dataset-1 and Dataset-2 is 98.53% and 90.10% respectively. The overall performance shows that the integrated approach is more suitable in detection and classification of Android malware.

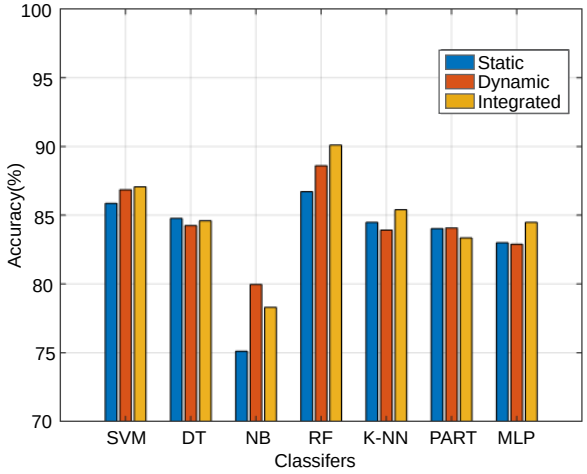



Fig. 16. Comparison of different classifiers based on accuracy using static, dynamic and integrated features for Dataset-2.

## V. Conclusion and Future Work

This paper presented a hybrid approach which extracts different types of features using static and dynamic malware analysis to detect and classify Android malware. We created our own two datasets for detection (dataset-1) and family classification (dataset-2) of Android malware. Both datasets consist of 352 static features and 323 dynamic features. These datasets are made publically available on GitHub and Kaggle with the aim to help researchers and anti-malware tool creators for enhancing or developing new techniques and tools for detecting and classifying Android malware. The significance of the datasets makes it appropriate to be used as benchmark to test new techniques. We employed the information gain feature selection algorithm to eliminate noisy and irrelevant features. Through this algorithm, we selected 110 and 47 static features in Dataset-1 and Dataset-2 respectively and 99 and 35 dynamic features in Dataset-1 and Dataset-2 respectively. The features with zero weights are not considered here. Various ML classifiers are applied to detect and identify Android malware. The experimental results indicate that the hybrid approach obtains better detection and classification performance as compared to the cases when static and dynamic features are considered alone. For dataset-1, RF provides the accuracy of 96.5% when only static features are considered and 97.01% when only dynamic features are considered. For dataset-2, RF provides accuracy of 86.72% when only static features are considered and 88.6% when only dynamic features are considered. RF provides the highest accuracy in the hybrid approach (when both static and dynamic features are integrated) for both Dataset-1 and Dataset-2 i.e. 98.53% and 90.1% respectively.

In real world scenario, the malware classification problem is a data imbalance problem as there exist more examples of benign applications as compared to the malicious ones. In future, we will focus on this issue while using deep learning and big data tools [57] to classify the Android malware applications.

## References


[1] StatistaReport. Accessed: December. 2019. [Online]. Available: http://www.statista.com/statistics/266488/forecast-of-mobile-appdownloads/.

[2] A. M. Memon, and A. Anwar, "Colluding apps: tomorrow's mobile malware threat," IEEE Security & Privacy, vol. 13 no. 6, pp. 77–81, 2015.

[3] Y. Zhou, and X. Jiang, "Dissecting Android malware: characterization and evolution," in IEEE Symposium in Security and Privacy, 2012, pp. 95–109.

[4] Future-Trends-of-Android-Malware-Growth. Accessed: December. 2019. [Online]. Available: https://www.researchgate.net/figure/Future-Trends-of-Android- Malware-Growth.

[5] McAfee Labs. (2018) Threat Predictions Report, McAfee Labs, Santa Clara, CA, USA.

[6] D. Barrera, H. G. Kayacik, P. C. V. Oorschot, and A. Somayaji, "A methodology for empirical analysis of permission-based security models and its application to Android," in Proc. of 17th ACM Conf. Computer and Communications Security, CCS 10, 2010, pp. 73–84.

[7] S. Singla, E. Gandotra, D. Bansal, and S. Sofat, "Detecting and classifying morphed malwares: A survey," International Journal of Computer Applications, vol. 122, no. 10, 2015.

[8] E. Gandotra, D. Bansal, and S. Sofat, "Malware analysis and classification: A survey," Journal of Information Security, vol. 5, no. 02, p. 56, 2014.

[9] CuckooDroid. Accessed: October. 2019. [Online]. Available: https://cuckoo-droid.readthedocs.io/en/latest/installation/.

[10] E. Gandotra, D. Bansal, and S. Sofat, "Malware intelligence: beyond malware analysis," International Journal of Advanced Intelligence Paradigms, vol. 13, no. 1-2, pp. 80-100, 2019.

[11] G. Suarez-Tangil, J. Tapiador, P. Peris-Lopez, and A. Ribagorda, "Evolution, detection and analysis of malware for smart devices," IEEE Communications Surveys & Tutorials, vol. 16, no. 2, pp. 961–987, 2013.

[12] S. Moghaddam, and M. Abbaspour, "Sensitivity analysis of static features for Android malware detection," in Electrical Engineering (ICEE), Tehran, Iran, 2014, pp. 920–924.

[13] Q. Li, and X. Li, "Android malware detection based on static analysis of characteristic tree," in Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), Xian, China, 2015, pp. 84-91.

[14] J. Li, L. Sun, Q. Yan, Z. Li, W. Srisaan, and Y. Heng, "Significant permission identification for machine-learning-based android malware detection," IEEE Transactions on Industrial Informatics, vol. 14, no. 7, pp. 3216-3225, 2018.

[15] H. J. Zhu, Z. H. You, Z. X. Zhu, W. L. Shi, X. Chen, and L. Cheng, "DroidDet: effective and robust detection of android malware using static analysis along with rotation forest model," Neurocomputing, vol. 272, pp. 638-646, 2018.

[16] S. Y. Yerima, and S. Sezer, "Droidfusion: A novel multilevel classifier fusion approach for android malware detection," IEEE transactions on cybernetic, vol. 49, no. 2, pp. 453-466, 2018.

[17] T. Kim, B. Kang, M. Rho, S. Sezer, and E. G. Im, "A multimodal deep learning method for Android malware detection using various features," IEEE Transactions on Information Forensics and Security, vol. 14, no. 3, pp. 773-788, 2018.

[18] A. Feizollah, N. B. Anuar, R. Salleh, G. S. Tangil, and S. Furnell, "Androdialysis: Analysis of android intent effectiveness in malware detection," Computers & Security, vol. 65, pp. 121-134, 2017.

[19] W. Wang, X. Wang, D. Feng, J. Liu, Z. Han, and X. Zhang, "Exploring permission-induced risk in android applications for malicious application detection," IEEE Transactions on Information Forensics and Security, vol. 9, no. 11, pp. 1869-1882, 2014.

[20] M. Dhalaria, E. Gandotra, and S. Saha, "Comparative Analysis of Ensemble Methods for Classification of Android Malicious Applications," in advances in Computing and Data Sciences, M. Singh, P. K. Gupta, V. Tyagi, J. Flusser, T. Oren and R. Kashyap, Eds. Singapore: Springer International Publishing, 2019, pp. 370-380.

[21] M. Dhalaria and E. Gandotra, "Convolutional Neural Network for Classification of Android Applications Represented as Grayscale Images," International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 8, no. 12S, pp. 835-843, 2019.

[22] H. Cai, N. Meng, B. Ryder, and D. Yao, "Droidcat: Effective android malware detection and categorization via app-level profiling," IEEE Transactions on Information Forensics and Security, vol. 14, no. 6, pp. 1455-1470, 2018.

[23] P. Feng, J. Ma, C. Sun, X. Xu, and Y. Ma, "A Novel Dynamic Android Malware Detection System With Ensemble Learning," IEEE Access, vol. 6, pp. 30996-31011, 2018.

[24] S. Das, Y. Liu, W. Zhang, and M. Chandramohan, "Semantics-based online malware detection: Towards efficient real-time protection against malware," IEEE transactions on information forensics and security, vol. 11, no. 2, pp. 289-302, 2015.

[25] W. Enck, P. Gilbert, S. Han, V. Tendulkar, B. Chun, L. P. Cox, J. Jung, P. McDaniel, and A. N. Sheth, "TaintDroid: an information-flow tracking system for real-time privacy monitoring on smartphones," ACM Transactions on Computer Systems (TOCS), vol. 32, no. 2, p. 5, 2014.

[26] L. Chen, M. Zhang, C. Y. Yang, and R. Sahita, "Semi-supervised classification for dynamic Android malware detection," arXiv preprint arXiv: 1704.05948, 2017.

[27] M. Zheng, M. Sun, and J. C. S. Lui, "DroidTrace: A ptrace based Android dynamic analysis system with forward execution capability," in international wireless communications and mobile computing conference (IWCMC), Nicosia, Cyprus, 2014, pp. 128-133.

[28] Z. Yuan, Y. Lu, and Y. Xue, "Droiddetector: android malware characterization and detection using deep learning," Tsinghua Science and Technology, vol. 21, no. 1, pp. 114-123, 2016.

[29] F. Tong, and Z. Yan, "A hybrid approach of mobile malware detection in Android," Journal of Parallel and Distributed computing, vol. 103, pp. 22-31, 2017.

[30] A. Martín, R. L. Cabrera, and D. Camacho, "Android malware detection through hybrid features fusion and ensemble classifiers: The AndroPyTool framework and the OmniDroid dataset," Information Fusion, vol. 52, pp. 128-142, 2019.

[31] T. Bläsing, L. Batyuk, A. D.Schmidt, S. A. Camtepe, and S. Albayrak, "An android application sandbox system for suspicious software detection," in 5th International Conference on Malicious and Unwanted Software,

Nancy, Lorraine, France, 2010, pp. 55-62.
[32] D. Arp, M. Spreitzenbarth, M. Hubner, H. Gascon, K. Rieck, and C. E. R. T. Siemens, "Drebin: Effective and explainable detection of android malware in your pocket," In Ndss, vol. 14, pp. 23-26, 2014.
[33] Virusshare. Accessed: March. 2019. [Online]. Available: https://virusshare.com/.
[34] APKMirror. Accessed: March. 2019. [Online]. Available: https://www.apkmirror.com/.
[35] Apkpure. Accessed: March. 2019. [Online]. Available: https://apkpure.com/.
[36] Avira. Accessed: April. 2019. [Online]. Available: https://www.avira.com/.
[37] W. Enck, D. Octeau, P. D. McDaniel, and S. Chaudhuri, "A study of android application security," In USENIX security symposium, vol. 2, p. 2, 2011.
[38] E. Gandotra, D. Bansal, and S. Sofat, "Tools & Techniques for Malware Analysis and Classification," International Journal of Next-Generation Computing, vol. 7, no. 3, 2016.
[39] Android4me: J2ME port of Google's Android (2011) https://code.google.com/p/android4me/downloads/list.
[40] Android Developers. Accessed: May. 2019. [Online]: Available: https://developer.android.com/guide/topics/manifest/permissionelement.
[41] Xposed module repository. Accessed: May. 2019. [Online]. Available: http://repo.xposed.info/module/de.robv.android.xposed.installer.
[42] S. Malik, and K. Khatter, "System call analysis of android malware families," Indian Journal of Science and Technology, vol. 9, no. 21, 2016.
[43] B. Chizi, and O. Maimon, "Dimension reduction and feature selection," in Data mining and knowledge discovery handbook, O. Maimon and L. Rokach, Eds. Boston MA: Springer, 2009, pp. 83-100.
[44] J. Han, J. Pei, and M. Kamber, "Data mining: concepts and techniques," Elsevier, 2011.
[45] G. Shakhnarovish, T. Darrell, and P. Indyk, "Nearest-neighbor methods in learning and vision," In MIT Press, 2005, p. 262.
[46] Keerthi, S. Sathiya, and E. G. Gilbert, "Convergence of a generalized SMO algorithm for SVM classifier design," Machine Learning, vol. 46, no. 1-3, pp. 351-360, 2002.
[47] A. Liaw, and M. Wiener, "Classification and regression by randomForest," R news, vol. 2, no. 3, 2002, pp. 18-22.
[48] J. R. Quinlan, "The Morgan Kaufmann Series in Machine Learning," San Mateo, 1993.
[49] P. Domingos, and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," Machine learning, vol. 29, no. 2-3, pp. 103-130, 1997.
[50] F. Eibe, and I. H. Witten, "Generating Accurate Rule Sets Without Global Optimization," In: Fifteenth International Conference on Machine Learning, 1998, pp. 144-151.
[51] S. B. Joo, S. E. Oh, T. Sim, H. Kim, C. H. Choi, H. Koo, and J. H. Mun, "Prediction of gait speed from plantar pressure using artificial neural networks," Expert Systems with Applications, vol. 41, no. 16, pp. 7398-7405, 2014.
[52] V. B. Semwal, K. Mondal, and G. C. Nandi, "Robust and accurate feature selection for humanoid push recovery and classification: deep learning approach," Neural Computing and Applications, vol. 28, no. 3, pp. 565-574, 2017.
[53] V. B. Semwal, M. Raj, and G. C. Nandi, "Biometric gait identification based on a multilayer perceptron," Robotics and Autonomous Systems vol. 65, pp. 65-75, 2015.
[54] V. B. Semwal, N. Gaud, and G. C. Nandi, "Human gait state prediction using cellular automata and classification using ELM," in machine intelligence and signal analysis, M.Tanveer and R. B. Pachori, Eds. Singapore: Springer, 2019, pp. 135-145.
[55] D. Gupta, and R. Rani, "Big Data Framework for Zero-Day Malware Detection," Cybernetics and Systems, vol. 49, no. 2, pp. 103-121, 2018.
[56] Scikit-Learn Machine Learning in Python. Accessed: June. 2019. [Online]. Available: https://scikit-learn.org/stable/.
[57] D. Gupta, and R. Rani, "A study of big data evolution and research challenges," Journal of Information Science, vol. 45, no. 3, pp. 322-340, 2019.

Meghna Dhalaria

Meghna Dhalaria is pursuing Ph.D. in Computer Science and Engineering Department at Jaypee University of Information and Technology, India. She has completed her Master's degree in Computer Science & Engineering from Thapar Institute of Engineering and Technology, Patiala. Her current research areas include the applications of machine learning and deep learning.

Ekta Gandotra

Ekta Gandotra is currently working as Assistant Professor in the Department of Computer Science and Engineering at Jaypee University of Information Technology, Waknaghat, India. She has around 12 years of teaching and research experience. She has completed her Ph.D. in Computer Science and Engineering from PEC University of Technology, Chandigarh, India. Her research areas include network & cyber security, malware threat profiling, cyber threat intelligence, machine learning and big data analytics.

# Video Data Compression by Progressive Iterative Approximation

M. J. Ebadi[1]*, A. Ebrahimi[2]

[1] Department of Mathematics, Chabahar Maritime University, Chabahar (Iran)
[2] Computer Geometry and Dynamical Systems Laboratory, Faculty of Mathematical Sciences, Yazd University, Yazd, (Iran)

Received 30 March 2020 | Accepted 12 November 2020 | Published 16 December 2020

unir
LA UNIVERSIDAD EN INTERNET

## Abstract

In the present paper, the B-spline curve is used for reducing the entropy of video data. We consider the color or luminance variations of a spatial position in a series of frames as input data points in Euclidean space R or R3. The progressive and iterative approximation (PIA) method is a direct and intuitive way of generating curve series of high and higher fitting accuracy. The video data points are approximated using progressive and iterative approximation for least square (LSPIA) fitting. The Lossless video data compression is done through storing the B-spline curve control points (CPs) and the difference between fitted and original video data. The proposed method is applied to two classes of synthetically produced and naturally recorded video sequences and makes a reduction in the entropy of both. However, this reduction is higher for syntactically created than those naturally produced. The comparative analysis of experiments on a variety of video sequences suggests that the entropy of output video data is much less than that of input video data.

## Keywords

B-spline Curve Fitting, Compression, Data Fitting, Least Square Fitting, Progressive and Iterative Approximation.

DOI: 10.9781/ijimai.2020.12.002

## I. Introduction

The technology of video compression has been a fundamental tool in video fields and multimedia communication for many years. The main objective of video compression is making a reduction in the volume of data by prospecting the correlations of video frames in such a way that a digital video file can be broadcast almost entirely over the network and stored on the computer disks. According to the required reconstruction, video compression techniques can be categorized into two large groups of lossy compression and lossless compression [1]–[3].

The increased use of high quality videos reveals the need for decreasing the volume of compressed video for transmission and storage, especially in social media networks. The lower the entropy of the data, the smaller the number of bits is required to encode them. Thus, this study aims to provide a practical procedure to reduce the entropy of the video data. Each color plane in the RGB space and subsequently RGB color image are respectively indicated using 8 bits/pixel and 24 bits/pixel. The inter and intra frame codings are exerted on the image sequences to decrease the temporal and spatial redundancy of the data in the image sequences.

The study of curve construction from a data point set is widely employed as a modeling instrument in many areas such as image processing, computer graphics, computer aided design (CAD), reverse engineering, object shape detection, and scientific visualization. According to its application, curves of implicit, parametric, and subdivision type are applied to data fitting. Converting data points into parametric curves including B-spline or Bézier curves is extremely desired in engineering applications. Most of the papers in the literature used motion estimation for video data compression. In more recent ones, the parametric curves such as the Bézier curve or natural cubic spline have been applied to compress video data into small storage space.

The Bézier curve is constructed by Bernstein basis that has limited flexibility. The degree of curve is directly relative to the number of control points (CPs). For a complicated shape and data, a large number of CPs may be required. To overcome this shortcoming and provide more flexibility and control, the B-spline curve is suggested as the generalization of the Bézier curve. In case the number of CPs is high, the use of lower degree parametric curves is possible. In order to prevent the additional cost of computations for solving a large linear equations system, the progressive iterative approximation (PIA) is used that is computationally efficient and simple to implement.

The PIA method is a direct and intuitive way of generating curve series of high and higher curve fitting accuracy. The PIA method refrains from solving a large linear equations system with an additional computational cost. The PIA technique begins with an initial curve and adjusts the curve CPs in an iterative process. Then, the resulted point cloud is interpolated and approximated by the limit curve. In this paper, we propose a technique for lossless video data compression making use of the B-spline curve The color or luminance variations of a spatial position in a series of frames are considered as data points in Euclidean space $\mathbb{R}$ or $\mathbb{R}^3$. The data points are approximated using progressive and iterative approximation for least square fitting (LSPIA). The proposed method reduces entropy and has efficient computational complexity.

* Corresponding author.
E-mail address: ebadi@cmu.ac.ir

In particular, our contributions are the following:

- We use the B-spline curve with remarkable flexibility to approximate the video data.
- The PIA method is applied to find the optimal CP of the B-spline curve with no need of solving a large linear equations system.
- Our method can be considered as a lossless video compression method that reduces the entropy of video data.

The organization of this paper is as follows. A brief summary of the related works is given in Section II. Section III provides a simple overview of LSPIA using the B-spline curve. The procedure of the proposed method using the B-spline curve to fit video data is explained in Section IV. Section V describes in detail the methodology adopted to design the video data compression. Section VI is dedicated to the study of the experimental results for various videos. A brief discussion on the proposed method in video compression is provided in Section VII. In the end, a conclusion is made in Section VIII.

## II. Review of Literature

In recent decades, multiple processes have been developed in the field of curve fitting by the use of Bézier and B-spline curves. Biswas [4] used the quadratic Bézier curve for compression of the grayscale images. Bézier curve has been used to capture the outline of planar generic images. An outline capturing technique was presented in [5] to estimate the appropriate location of CPs by the utilization of the cubic Bézier curve properties. In [6], a method is designed to capture the outline of 2D shapes using the cubic Bézier curve with the emphasis given to local control of data points rather than the global error of square fitting. A novel outline capturing scheme for 2D shapes was introduced in [7] based on the Nelder-Mead simplex method.

In [8], the L-BFGS optimization is exerted on data points to which B-spline curve is fitted. Ebrahimi and Loghmani [9] used approximation BFGS methods to make optimization of the foot and CPs of the B-spline curve. The complexity per step in [9] is $O(n)$, requiring only $O(n)$ memory allocations. In [10], a practical approach to curve fitting is presented for the specification of the initial B-spline curve which is near to the target curve. A length parameter is presented by this method which allows adjustment to the number of CPs. This makes the initial B-spline curve more precise. The scaled BFGS algorithm is then employed for simultaneous optimization of control and foot points.

Lin et al. [11] introduced the phrase "progressive iterative approximation" in 2005. The standard PIA procedure is not feasible for curve fitting with plenty of the data points when control and data points are equal in number [12]. Delgado and Pena [13] proved that the normalized B-basis is a totally positive basis with the fastest convergence rate. A local PIA format is designed in [14] and showed the convergence of the local format for the normalized totally positive based blending curve. An approach is proposed for weighted PIA of data points using normalized totally positive basis in [15] with a faster convergence rate. In [16], an extended PIA is introduced in which the number of given data points with storage requirement $O(n)$ is higher than the number of CPs, where n is the number of the CPs. An adaptive data point fitting based on the PIA is proposed in [17]. Zhang et al. [18] developed a progressive T-spline method of fitting large-scale datasets such as images of high precision. Deng and Lin [19] introduced the LSPIA where the number of data points is more than that of CPs. LSPIA provides a set of fitting curves making adjustments of the CPs and leading to the given data points through least square (LS) fitting as the final curve. Ebrahimi and Loghmani [20] presented the composite iterative method for LSPIA with a fast convergence rate. This method constructs a series of matrices applied to the adjusting vector on the base of the Schulz iterative method. A comprehensive survey on PIA methods has been provided in [21].

Motion Estimation (ME) is the most popular in removing the temporal redundancy in video compression that can be arranged into pixel and block motion estimations [22]. The motion vector in pixel motion estimation is computed for every pixel in the frame. The block motion estimation method divides frames into blocks and then the motion vector is computed for every block. In the interframe coding method, block motion estimation plays a key role in reducing temporal redundancy in the image sequence. A block-matching approach can be developed to modify the coding efficiency and video quality. In the past three decades, some improvements have been made in motion estimation techniques such as pelrecursive methods, optical flow, block matching algorithms, and parametric-based models [23].

By making some attempts, quick application and simple comprehension of block matching algorithms make them fundamental methods of motion estimation in video compression. The full search algorithm (FSA) is the easiest method in the block matching algorithms that has high computational cost. To accelerate the search procedure and decrease the computational complexity, several fast block matching algorithms, such as diamond search (DS) [24], hexagon search (HS) [25], three step search (TSS) [26], [27], four step search (FSS) [28] have been proposed.

To perform a method of fast motion estimation, Koga et al. [29] introduced TSS as a primary attempt. Compared to the full search, the TSS method has a less computational cost in terms of average search point and mean absolute difference.

The computational cost TSS method is less in average search point and mean absolute difference as compared to the full search method. The modified TSS algorithm is presented in [27] for weighted finite automata coding and block matching motion estimation methods to reduce the encoding time.

Video and image compression using parametric curves explored by many authors. Fu et al. [30] has been explored a video object encoding method pursuant to the data fitting trajectory of video object moving edges pixels that is suitable for the slowly moving video/ video data. The cubic spline interpolation is used in [31] to modify medical image compression for medicine applications. The cubic convolution spline interpolation is proposed on the basis of the LSs method to compress the image data in [32]. In [33], based on the natural cubic spline and parametric line fitting, a method for lossy compression is presented in order for compression of digital video data in the temporal dimension. The linear Bézier curve is used in [34] for the approximation of temporal video sequence in Euclidean space. Khan [35] has proposed an algorithm for lossless video compression which was based on the quadratic Bézier curve and least square technique.

## III. The Progressive and Iterative Approximation for Least Square Fitting (LSPIA)

Here, we first formulate the blending curve and review the LSPIA (readers are referred to [19] for details).

A nonnegative basis $\{N_i(t); i = 0, 1, ..., n\}$ defined on a set $I$ with $\sum_{i=1}^{n} N_i(t) = 1$ for all $t \in I$ is taken as a blending basis.

A totally positive blending basis is defined as normalized totally positive (NTP) basis. Let $\{N_i(t); i = 1, 2, \cdots, n\}$ be an NTP blending basis. Then, assuming a sequence of the CPs $\{P_i\}_{i=0}^{n}$ in $\mathbb{R}$ or $\mathbb{R}^3$, a blending curve as

$$C(t) = \sum_{i=0}^{n} P_i N_i(t) \tag{1}$$

can be considered. Suppose that $\{q_j; j = 1, 2, \cdots, m\}$ is an ordered data point sequence on a target curve to be fitted and

$$\Gamma = \{0 = t_0 < t_1 < \cdots < t_m = 1\}$$

is the location parameters of $\{q_j; j = 1, 2, \cdots, m\}$. Taking $\{P_i^0\}_{i=0}^n$ form $\{q_j; j = 1, 2, \cdots, m\}$ similar to the CPs, the initial blending curve $C^0(t)$ is defined:

$$C^0(t) = \sum_{i=0}^{n} P_i^0 N_i(t), \quad t \in [t_0, t_m] \tag{2}$$

The $(m+1) \times (n+1)$ collocation matrix of the NTP blending basis $\{N_i(t); i = 1, 2, \cdots, n\}$ on $\Gamma$ is

$$A = \begin{bmatrix} N_0(t_0) & N_1(t_0) & \ldots & N_n(t_0) \\ N_0(t_1) & N_1(t_1) & \ldots & N_n(t_1) \\ \ldots & \ldots & \ldots & \ldots \\ N_0(t_m) & N_1(t_m) & \ldots & N_n(t_m) \end{bmatrix} \tag{3}$$

At the beginning of the iteration, let

$$\delta_j^0 = q_j - C^0(t_j), \ \ j = 0, 1, \cdots, m \tag{4}$$

$$\Delta_i^0 = \mu \textstyle\sum_{j=0}^{m} N_i(t_j)\delta_j^0, \ i = 0, 1, ..., n \tag{5}$$

where $\mu$ is a non zero real scalar and

$$0 < \mu < \frac{2}{\lambda_0}$$

where $\lambda_0$ is the largest eigenvalue of $A^T A$. By the movement of the CPs $P_i^0$ along the regulating vector $\Delta_i^0$, i.e.

$$P_i^1 = P_i^0 + \Delta_i^0, \ \ i = 0, 1, ..., n \tag{6}$$

and the new curve,

$$C^1(t) = \sum_{i=0}^{n} P_i^1 N_i(t), \ \ t \in [t_0, t_m] \tag{7}$$

Similarly, obtaining the $k$-th blending curve $C^k$ after the $k$-th iteration, we suppose

$$\delta_j^k = q_j - C^k(t_j), \quad j = 0, 1, \cdots, m \tag{8}$$

$$\Delta_i^k = \mu \textstyle\sum_{j=0}^{m} N_i(t_j)\delta_j^k, \quad i = 0, 1, ..., n \tag{9}$$

we can generate the $(k+1)$-th blending curve as follows

$$C^{k+1}(t) = \sum_{i=0}^{n} P_i^{k+1} N_i(t), \quad t \in [t_0, t_m] \tag{10}$$

The mentioned iterative process produces a curve sequence $\{C^k(t), k = 0, 1, ...\}$ whose limit is the LS fitting curve of the original data points $\{q_j\}_{j=0}^m$ [19].

The initial situation of the CPs ($\{P_i^0\}_{i=0}^n$) may be selected as

$$P_0^0 = q_0, \ P_n^0 = q_m$$

$$P_i^0 = q_{g(i)}, \ i = 1, 2, ..., n-1$$

where $g(i) = \left[\frac{(m+1)i}{n}\right]$. In addition, we adopt the uniform parametrization to assign the parameters $\{t_j\}_{j=0}^m$ for $\{q_j\}_{j=0}^m$.

In this study, having numerical computation stability and extensive use in image processing, we wxamine the LSPIA by B-spline curve. Having B-spline basis functions, B-spline is a blending curve. Let $\{P_i\}_{i=0}^n$ be $n+1$ CPs and $N_{i,r}(t)$ be the B-spline basis functions of degree $r$ (or order $r+1$) defined on a given nondecreasing real-number knot vector $U = \{u_0, u_1, \cdots, u_{n+r+1}\}$, then a B-spline curve of degree $r$ will be as follows

$$C(t) = \sum_{i=0}^{n} P_i N_{i,r}(t) \tag{11}$$

where the B-spline basis functions $N_{i,r}(t)$ is defined recursively by the Boor formula

$$N_{i,0}(t) = \begin{cases} 1, & u_i \leqslant t \leqslant u_{i+1} \\ 0, & \text{otherwise} \end{cases}$$

$$N_{i,r}(t) = \frac{t - u_i}{u_{i+r-1} - u_i} N_{i,r-1}(t) + \frac{u_{i+r} - t}{u_{i+r} - u_{i+1}} N_{i+1,r-1}(t), r \geqslant 1$$

The proposed method works for any number of CPs and any degree of B-spline curve but from our experimental results, we notice that $r = 3$ and $n = 7$ are appropiate for degree of curve and number of CPs respectively. Further, the cubic B-spline basis are constructed on the knot vectors

$$u_0 = u_1 = ... = u_3 = 0$$

$$u_{j+3} = (1 - \beta)t_{i-1} + \beta t_i, \ \ j = 1, ..., n-3$$

$$u_{n+1} = ... = u_{n+4} = 1$$

where $i = [jd], \beta = jd - i, d = \frac{m+1}{n-2}$. According to the B-spline curve definition, it is clear that the properties of the B-spline basis function are passed t the B-spline curve. These properties are as follows:

- Partition of unity
- Affine invariance
- Convex hull property
- Local control
- Multiple knots

## IV. Video Data Fitting With LSPIA

In this section, the process of the video data fitting using LSPIA is presented. Let a video include a sequence of m frames, and each frame possesses $W \times H$ pixels, where $H$ and $W$ respectively are the height and width of video frames. The value of each pixel in a frame is a data point in Euclidean space $\mathbb{R}^1$ or $\mathbb{R}^3$ for luminance or 3-D RGB, respectively. The temporal data of a spatial location

$$(x, y), 1 \leqslant x \leqslant W, 1 \leqslant y \leqslant H$$

in $m$ frames are $\{q_1, q_2, \cdots, q_m\}$, i. e., $\{q_j = I_j\}_{j=0}^m$ for luminance or $\{q_j = (r_j, g_i, b_i)\}_{j=0}^m$ for 3-D RGB. Then, we approximate the $m$ values of each spatial location $\{q_1, q_2, \cdots, q_m\}$ by the LSPIA method. Fig. 1 illustrates the RGB variation of a spatial position (50, 50) in 96 Mobile and Calendar video sequence frames.

The video data from each spatial location in an sequence of frames (input data) is approximated with much less number of control points (output data) of the B-spline curve. This process is separately used to intensify RGB variations in the temporal dimension of each spatial position. The luminance values of a spatial position (50, 50) in 96 Foreman video sequence frames are fitted using a cubic B-spline basis and LSPIA in Fig. 2.

## V. Methodology

The main purpose of the proposed method is video data compression by reducing the entropy of output data. A smaller number of bits is needed to store the video data with lower entropy. In the first step, the color or luminance variations of a spatial position in series of frames are considered the input data in Euclidean space $\mathbb{R}$ or $\mathbb{R}^3$. In the next step, we use the B-spline curve and approximate the input video data with a considerably smaller number of CPs. In addition, LSPIA fits the input video data with low approximation errors and without solving a system of equations. In this step of the work, we need to store only the CPs of the B-spline curve to approximate the input data. In the

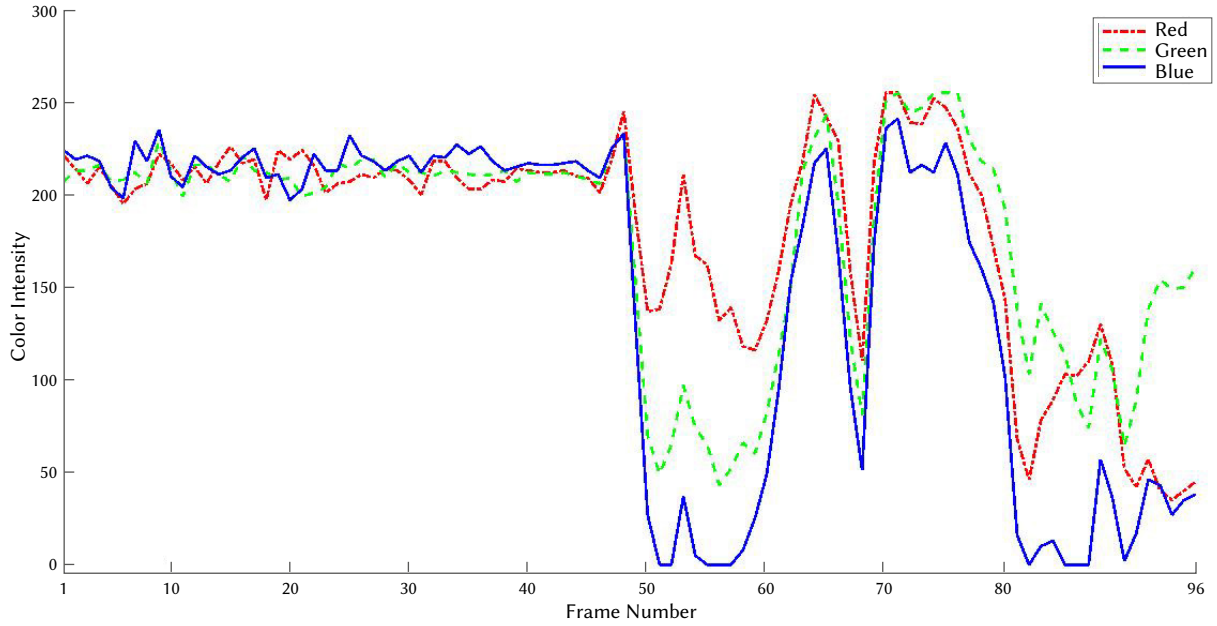



Fig. 1. The RGB variation of a spatial position (50, 50) in 96 Mobile and Calendar video sequence frames.

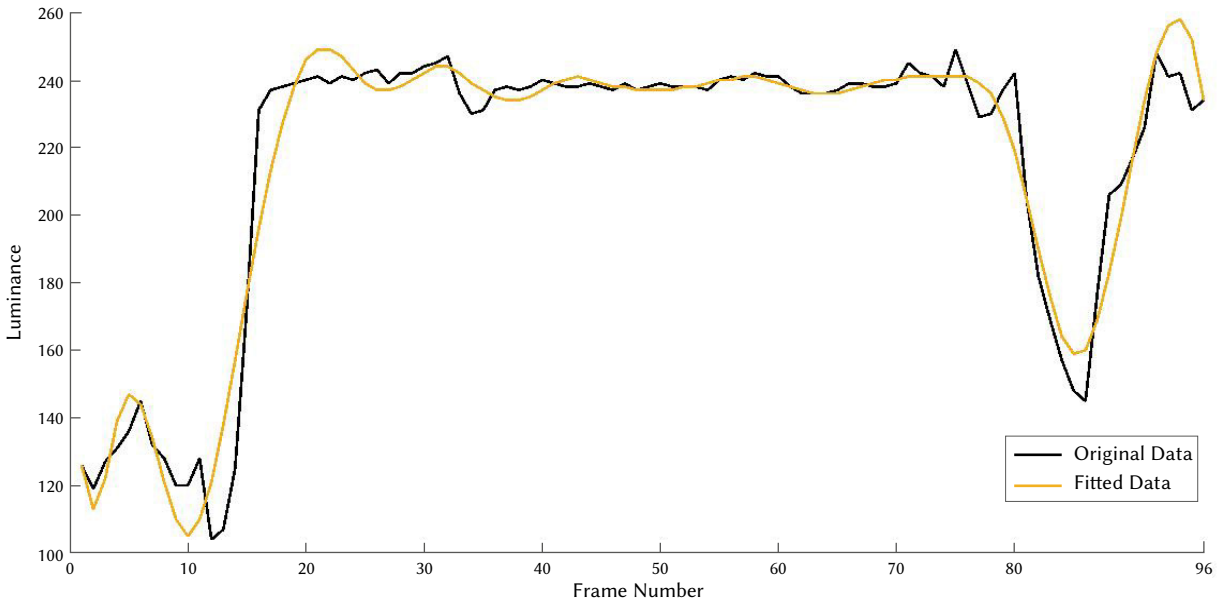



Fig. 2. B-spline curve fitting to the luminance values of a spatial position (50, 50) in 96 Foreman video sequence frames using LSPIA.

final step, the difference between primary and approximated video frames (DF) are also stored for lossless video compression. According to the curve fitting method used in this work, the difference between primary and cubic B-spline approximated data has limited values in the short-range in comparison to primary values in the primary video sequence. Therefore, the entropy of CP and DF in the proposed method is far less than that of the primary video sequence. It is worth mentioning that our method can be used for lossy video compression. The basic foundation of our method is explained in Algorithm 1.

We use the CPs of the B-spline curve to create the approximated video frames and then add the frame difference (FD) to reconstruct the original video. In contrast to the most existing methods that use the neighbor's pixels to reduce spatial redundancy, our method merely uses temporal redundancy.

**Algorithm 1**. Video data compression using LSPIA

**Input**: A video includes a sequence of $m$ frames, and each frame possesses $W \times H$ pixels;

**Output**: The CPs of B-spline curve and the difference between primary and approximated video frames (DF);

**for** $i = 1$ **to** $W$ **do**

  **for** $j = 1$ **to** $H$ **do**

    Consider the data of spatial location $(i, j)$ in $m$ frames $\{q_k\}_{k=1}^{m}$;

    Approximate $\{q_k\}_{k=1}^{m}$ using LSPIA;

    Store the CPs of B-spline curve;

    Store the difference between $\{q_k\}_{k=1}^{m}$ and B-spline curve (DF);

  **end for**

**end for**

a) Mobile and calendar sequence

b) Dinosaur sequence

c) Cloud sequence

d) Foreman sequence

e) Hall and monitor sequence

Fig. 3. One of the frames of video sequences.

## VI. Experiments and Results

The introduced technique described in the previous section has been applied to some synthetically produced and naturally recorded video sequences and its results have been compared with those obtained with the method in [35]. We compare our method with the technique proposed by khan [35] because it used the quadratic Bézier curve and entropy criterion.

According to the required reconstruction, the methods of video compression can be classified into two groups of lossless compression methods, in which the output video is identical to input video, and lossy compression methods, with generally provide much higher compression in which the output video is different from the input video.

Some innovative improvements have been recently made to lossy video compression to which interested readers can refer [27] and the references therein. The introduced method in this study is a lossless video compression and hence instead of PSNR, we use the entropy criterion to evaluate the efficiency of the compression method. The entropy is a scale of the required mean number of binary symbols for coding the source output. Encoding source output with the bit mean number equal to the source entropy is indicative of a desired lossless compression method.

Suppose a source (frame) of information has $M$ symbols (pixel values) with individual probability $P_i$ and

$$\sum_{i=1}^{M} P_i = 1$$

The entropy of a single video frame can be defined by:

$$H = -\sum_{i=1}^{M} P_i \log P_i \tag{12}$$

We calculate the entropy of video using the mean entropy of all frames that construct the video sequence.

To evaluate the proposed method efficiency, five standard video sequences of different resolutions with sufficient complexity are selected for the simulation as listed in Table I and one of the frames of each input video sequence is represented in Fig. 3.

TABLE I. Schematic of the Test Video Sequences

| Test video sequences | Format | Resolution | Frames |
|---|---|---|---|
| Mobile and calendar | RGB | 352 × 240 | 96 |
| Dinosaur | RGB | 352 × 288 | 96 |
| Cloud | RGB | 352 × 240 | 96 |
| Foreman | Luminance | 352 × 240 | 96 |
| Hall and monitor | Luminance | 352 × 28896 | 96 |

The entropy of the videos is simply calculated using Equation (12). The output data in our method for computing entropy consists of the CPs and the difference between primary and approximated video frames (DF). The output data produced by algorithm [35] need to be stored and used in computing entropy which includes: (1) the end CPs of Bézier curve, (2) the middle CPs of Bézier curve, and (3) the difference between the quadratic Bézier approximated and original video sequences.

Table II compares the introduced method with algorithm [35] in terms of entropy. It can be seen that the significantly lower entropy is produced by the proposed method than those generated by [35].

TABLE II. Performance Comparison in Terms of Entropy

| Video name | Original video | Method [35] | Our method |
|---|---|---|---|
| Mobile and calendar | 7.627 | 6.653 | 6.431 |
| Dinosaur | 7.163 | 2.736 | 2.334 |
| Cloud | 7.567 | 4.032 | 3.849 |
| Foreman | 7.228 | 5.124 | 4.576 |
| Hall and monitor | 7.233 | 3.918 | 3.243 |

## VII. Discussion

The videos tested in the previous section are classified into two groups: (a) naturally recorded video sequences; and (b) synthetically created video sequences. Among them Hall, Mobile and Foreman video sequences are naturally recorded, while Cloud and Dinosaur video sequences are synthetically produced. Mobile and Calendar, Dinosaur and Cloud video sequences have RGB components while Foreman and Hall have a single component of luminance. Although our proposed method makes a decrease in the entropy of both classes of video sequences, the entropy of naturally recorded is more decreased than that of synthetically produced. In fact, the proposed algorithm performs significantly better for the synthetically created video sequences. It can be justified that the synthetically created video sequences have less temporal fluctuations and can be approximated with a small number of CPs. The number and degree of CPs in the B-spline curve are two factors that must be determined in our method.

The causes for the performance of our method are as follows:

1. Instead of the Bézier curve, the B-spline curve is used in our proposed method which has better interactive flexibility and local control property. Also, the number of CPs can be changed with no need of changing the degree of the B-spline curve. Hence, the introduced method creates a better approximation with desirable precision.
2. The input data in [35] are divided into segments based on the breakpoints and each segment is then approximated by a quadratic Bézier curve. This is while our method fits the input data without segmentation using a B-spline curve. This makes simpler computations for the method proposed compared to the method [35].
3. The LSPIA method used in this study approximates the video data with low fitting errors and without solving a system of equations.
4. The reduction in entropy is higher for synthetically produced than naturally recorded video sequences.
5. In comparison with block level fitting, the pixel level fitting provides more control over accuracy.

The other PIA methods such as composite iterative method with fast convergence rate [20] can be further used to find the optimal CP of the B-spline curve. The weighted parametric curves like the NURBS curve can be applied instead of the B-spline curve. This is a topic of interest for our future work. The authors plan to use the proposed method in H.264 coding that is a modern video compression method with lossless macro-block coding features.

## VIII. Conclusion

A practical method for lossless video compression with a B-spline curve has been introduced. The purpose of our method was to fit the data obtained from the color or luminance variations of a spatial position in series of frames. The LSPIA found the optimal CPs and approximated the input data. The introduced method can be used for 3-D color spaces such as *RGB*, $Y C_b C_r$ or *HSV*. The experimental results demonstrated an easier implementation of our proposed algorithm and substantially reduced entropy of video sequences. The superiority of our study lies behind the fact that it causes a reduction in the entropy of all video sequences, particularly the synthetically created ones.

## References


[1] K. Sayood, *Introduction to data compression*. Morgan Kaufmann, 2017.

[2] E. V. Pérez, M. Sánchez, R. G. Crespo, *et al.*, "A system to generate signwriting for video tracks enhancing accessibility of deaf people.," *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 4, no. 6, 2017.

[3] R. C. Joshi, A. G. Singh, M. Joshi, S. Mathur, "A low cost and computationally efficient approach for occlusion handling in video surveillance systems," *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 5, no. 7, pp. 28–38, 2019.

[4] S. Biswas, "One-dimensional b–b polynomial and hilbert scan for graylevel image coding," *Pattern Recognition*, vol. 37, no. 4, pp. 789–800, 2004.

[5] M. Sarfraz, A. Masood, "Capturing outlines of planar images using bézier cubics," *Computers & Graphics*, vol. 31, no. 5, pp. 719–729, 2007.

[6] G. Barid Loghmani, A. Ebrahimi, M. Sarfraz, "Capturing outlines of planar generic images by simultaneous curve fitting and subdivision," *Journal of AI and Data Mining*, 2019.

[7] A. Ebrahimi, G. Loghmani, M. Sarfraz, "Capturing outlines of generic shapes with cubic bézier curves using the nelder–mead simplex method," *Iranian Journal of Numerical Analysis and Optimization*, vol. 9, no. 2, pp. 103–121, 2019.

[8] W. Zheng, P. Bo, Y. Liu, W. Wang, "Fast b-spline curve fitting by l-bfgs," *Computer Aided Geometric Design*, vol. 29, no. 7, pp. 448–462, 2012.

[9] A. Ebrahimi, G. B. Loghmani, "B-spline curve fitting by diagonal approximation bfgs methods," *Iranian Journal of Science and Technology, Transactions A: Science*, vol. 43, no. 3, pp. 947–958, 2019.

[10] A. Ebrahimi, G. B. Loghmani, "Shape modeling based on specifying the initial b-spline curve and scaled bfgs optimization method," *Multimedia Tools and Applications*, vol. 77, no. 23, pp. 30331–30351, 2018.

[11] H. W. Lin, H. J. Bao, G. J. Wang, "Totally positive bases and progressive iteration approximation," *Computers & Mathematics with Applications*, vol. 50, no. 3-4, pp. 575–586, 2005.

[12] Y. Kineri, M. Wang, H. Lin, T. Maekawa, "B-spline surface fitting by iterative geometric interpolation/approximation algorithms," *Computer-Aided Design*, vol. 44, no. 7, pp. 697–708, 2012.

[13] J. Delgado, J. M. Peña, "Progressive iterative approximation and bases with the fastest convergence rates," *Computer Aided Geometric Design*, vol. 24, no. 1, pp. 10–18, 2007.

[14] H. Lin, "Local progressive-iterative approximation format for blending curves and patches," *Computer Aided Geometric Design*, vol. 27, no. 4, pp. 322–339, 2010.

[15] L. Lu, "Weighted progressive iteration approximation and convergence analysis," *Computer Aided Geometric Design*, vol. 27, no. 2, pp. 129–137, 2010.

[16] H. Lin, Z. Zhang, "An extended iterative format for the progressive-iteration approximation," *Computers & Graphics*, vol. 35, no. 5, pp. 967–975, 2011.

[17] H. Lin, "Adaptive data fitting by the progressive-iterative approximation," *Computer aided geometric design*, vol. 29, no. 7, pp. 463–473, 2012.

[18] H. Lin, Z. Zhang, "An efficient method for fitting large data sets using t-splines," *SIAM Journal on Scientific Computing*, vol. 35, no. 6, pp. A3052–A3068, 2013.

[19] C. Deng, H. Lin, "Progressive and iterative approximation for least squares b-spline curve and surface fitting," *Computer-Aided Design*, vol. 47, pp. 32–44, 2014.

[20] A. Ebrahimi, G. B. Loghmani, "A composite iterative procedure with fast convergence rate for the progressive-iteration approximation of curves," *Journal of Computational and Applied Mathematics*, vol. 359, pp. 1–15, 2019.

[21] H. Lin, T. Maekawa, C. Deng, "Survey on geometric iterative methods and their applications," *Computer-Aided Design*, vol. 95, pp. 40–51, 2018.

[22] H. Amirpour, M. Ghanbari, A. Pinheiro, M. Pereira, "Motion estimation with chessboard pattern prediction strategy," *Multimedia Tools and Applications*, vol. 78, no. 15, pp. 21785–21804, 2019.

[23] I. Chakrabarti, K. N. S. Batta, S. K. Chatterjee, *Motion Estimation for Video Coding*. Springer, 2015.

[24] S. Zhu, K. K. Ma, "A new diamond search algorithm for fast block-matching motion estimation," *IEEE transactions on Image Processing*, vol. 9, no. 2, pp. 287–290, 2000.

[25] C. Zhu, X. Lin, L. P. Chau, "Hexagon-based search pattern for fast block motion estimation," *IEEE transactions on circuits and systems for video technology*, vol. 12, no. 5, pp. 349–355, 2002.

[26] X. Jing, L. P. Chau, "An efficient three-step search algorithm for block

motion estimation," *IEEE transactions on multimedia*, vol. 6, no. 3, pp. 435–438, 2004.

[27] S. D. Kamble, N. V. Thakur, P. R. Bajaj, "Modified three-step search block matching motion estimation and weighted finite automata based fractal video compression.," *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 4, no. 4, pp. 27–39, 2017.

[28] L. M. Po, W. C. Ma, "A novel four-step search algorithm for fast block motion estimation," *IEEE transactions on circuits and systems for video technology*, vol. 6, no. 3, pp. 313–317, 1996.

[29] T. Koga, "Motion compensated interframe coding for video-conferencing," in *Proc. Nat. Telecommun. Conf.*, 1981, pp. G5–3.

[30] X. Fu, D. Liang, D. Wang, "A new video compression algorithm for very low bandwidth using curve fitting method," in *International Conference on Advances in Visual Information Systems*, 2007, pp. 223–229, Springer.

[31] T. K. Truong, S. H. Chen, T. C. Lin, "Medical image compression using cubic spline interpolation with bit-plane compensation," in *Medical Imaging 2007: PACS and Imaging Informatics*, vol. 6516, 2007, p. 65160D, International Society for Optics and Photonics.

[32] T. K. Truong, L. J. Wang, I. S. Reed, W. S. Hsieh, "Image data compression using cubic convolution spline interpolation," *IEEE Transactions on Image Processing*, vol. 9, no. 11, pp. 1988–1995, 2000.

[33] M. A. Khan, Y. Ohno, "Compression of video data using parametric line and natural cubic spline block level approximation," *IEICE transactions on information and systems*, vol. 90, no. 5, pp. 844–850, 2007.

[34] M. A. Khan, "An automated algorithm for approximation of temporal video data using linear b'ezier fitting," *The International Journal of Multimedia and Its Applications*, vol. 2, no. 2, pp. 81–94, 2010.

[35] M. A. Khan, "A new method for video data compression by quadratic bézier curve fitting," *Signal, Image and Video Processing*, vol. 6, no. 1, pp. 19–24, 2012.

M. J. Ebadi

M. J. Ebadi received his BSc degree in Applied Mathematics from Shahid Bahonar University of Kerman, Iran in 2003 and his MSc degree in Applied Mathematics from S&B University, Iran in 2006. In the same year, he joined the Department of Mathematics at Chabahar Maritime University, Iran, as a faculty member. In 2018, he received his Ph.D. in Applied Mathematics from Yazd University, Iran. He is a member of editorial board of two international reputed journals and a reviewer of more than 20 international reputed journals. He has published several international papers in high rank journals. Currently, he is an Assistant Professor in Applied Mathematics at the Department of Mathematics, Chabahar Maritime University, Iran. His research interests include numerical optimization, deep learning, neural networks, numerical analysis, optimal control, fuzzy optimization, numerical optimization, EEG signals classification, and image processing.

A. Ebrahimi

A. Ebrahimi is a researcher in computeraided geometric design at the computer geometry and dynamical systems laboratory of Yazd University, Yazd, Iran. He received his Ph.D. degree in Applied Mathematics from Yazd University in 2019. His research interest includes geometric modeling and image processing.

# Does a presentation Media Influence the Evaluation of Consumer Products? A Comparative Study to Evaluate Virtual Reality, Virtual Reality with Passive Haptics and a Real Setting

Julia Galán[1], Carlos García-García[1], Francisco Felip[1]*, Manuel Contero[2]

[1] Universitat Jaume I, Castellón (Spain)
[2] Universitat Politècnica de València, Valencia (Spain)

Received 17 Sep 2020 | Accepted 4 January 2021 | Published 11 January 2021

## Abstract

Technologies based on image offer a high potential to present consumers with products by focusing on their visual characteristics, but lack the capacity to physically interact with an object, which can compromise how consumer products are evaluated. The present study aims to analyse the influence of different presentation media on how users perceive the product by comparing the evaluation of a piece of furniture made by a sample of 203 users, which was presented in three different settings: a real setting (R), a Virtual Reality setting (VR) and a Virtual Reality with Passive Haptics setting (VRPH). To evaluate the product in the different settings, a semantic differential scale was built that comprised 12 bipolar pairs of adjectives. To study the results, the descriptive statistics for the semantic differential scales were analysed, a study about the frequency of repetition was conducted of each evaluation, a Kruskal-Wallis test was conducted and Dunn's post hoc tests were performed. The results showed that the presentation media of a piece of furniture influenced the evaluation of how users perceived it. These results also revealed that the haptic interaction with a product influenced how users perceived it compared to an exclusively visual interaction.

## Keywords

Consumer Products, Design, Perception And Psychophysics, Product Evaluation, Virtual Reality.

DOI: 10.9781/ijimai.2021.01.001

## I. Introduction

As e-retailing is becoming increasingly more frequent, the traditional ways of presenting products are being gradually replaced with digital media based on image use [1]–[3]. These media are generally employed to present collections of images, videos or motion graphics to show different product characteristics on some type of screen. This gives users an idea about what the product is like and they can evaluate its suitability to make informed purchase decisions [4].

Nonetheless, some product characteristics are not easy to evaluate with such presentation media owing to their bidimensional and visual limitations. Such is the case of object's real volume, tridimensionality of all its components or surface finishing. Nor is it easy to evaluate characteristics related to sensorial processes other than visual ones, such as comfort, tactile texture or weight.

Many studies are available about the main factors that influence consumers when they make purchasing decisions, and the following stand out: being familiar with the product [5], the setting in which the object is put on show [6], [7], the product's aesthetics [8] or the influence of how a product is presented in digital media [3]. Holbrook and Hirschman [9] point out that consumer behaviour is generally studied from a rational choice perspective, where less attention is paid to a visual experience that takes into account entertainment activities, sensorial pleasures or emotional responses. Accordingly, another work [10] studies the influence of consumer attitudes when making purchasing decisions, defining attitudes as those lasting evaluations made by the user of an object, a theme or a person [11].

Jordan [12] identifies 3 hierarchical levels as regards of consumer requirements; 1: Functionality, 2: Usability, 3: Pleasure; and also identifies four pleasures that people may seek and that products may bring about [13]: physio-pleasure (pleasures deriving from sensorial organs, like touch, taste or smell), socio-pleasure (pleasures deriving from relationships with others, e.g., friends loved ones or people who hold similar ideas), psycho-pleasure (pleasures related to people's cognitive and emotional reactions) and ideo-pleasure (pleasures related to people's values, such as a product's aesthetics, or the values it represents for some reason, like social or environmental responsibility).

Generally speaking, the intention is for the product's presentation mode to be as truthful as possible about the product's attributes and qualities, which directly affect how users perceive it. Wu et al. [14] defend the notion that the quality of images impacts how a product is understood. Other authors [15] state that the size, quality and

* Corresponding author.
E-mail address: ffelip@uji.es

movement of a product's image influence how consumers perceive its degree of usability, which conditions their decision making.

Artacho-Ramirez et al. [16] found a significant influence of the mode of representation on the product perception although differences were less numerous than expected (only 3 out of 11 semantic axes employed for the evaluation). It is worth mentioning that differences decreased as more sophisticated visualizations media were employed such as a navigable 3D model.

Naderi et al. [17] studied the combined effects of product design and environment congruence on consumers' aesthetic, affective and behavioral responses. The experimental stimuli used in their study, were presented in a 3D simulation environment using a large TV, and a stereoscopic virtual reality headset. They found that while most of the findings were similar across the two presentation media, there were a few discrepancies attributed to the use of different navigation methods and much closer experience to reality, for the VR headset.

The Augmented Reality (AR), Virtual Reality (VR) and Mixed Reality (MR) technologies have long been changing the way products are presented to consumers. These technologies allow us to go beyond 2D screen limitations by offering consumers a more immersive and interactive experience. Depending on the technology employed and its limitations, users can be immersed in a 3D virtual world, move around it, and can even interact with some elements represented in it using different devices.

It is interesting to study the use of different technologies to virtually present products to consumers without them having to travel to a physical point of sales and which also guarantees the correct visualization of their 3D characteristics. Verhagen et al. [18] stresses that using Virtual Mirrors improves how products are perceived in relation to using 360 spin or images. Suh and Lee [19] point out that using VR increases consumer knowledge and their purchase intention. Grewal et al. [20] point out that the greater immersion and interactivity provided by the product's VR representations allow more information to be obtained about the product and improve the user's experience.

Nonetheless, completely virtual technologies that offer a high potential to present consumers with products by focusing on their visual characteristics lack the capacity to provide physical interaction with an object. This limitation can compromise how consumer products are evaluated in completely virtual settings [21]. Therefore, different research works have studied consumers' need to touch a product to make a purchasing decision [22], [23] [24]. In order to overcome this barrier, some physical objects can be included in a controlled virtual setting so that users can live a more immersed experience in the VR setting by interacting with and feeling some virtual objects they see. We refer to such settings as Virtual Reality with Passive Haptics (VRPH).

Interactions with VRPH settings can provide advantages of coming into haptic contact with the object, along with the possibility of interacting and modifying the virtual setting. This allows the textures, colours, surface finishings or materials of the presented product to be altered in real time so that the range of physical products needed to offer users the whole brand's physical showroom catalogue can be reduced. Instead only making one product physically available would be necessary to provide its shape, tactile texture, materials and real reliefs, regardless of visual finishing touches, so that users could perceive all its other characteristics (colour, pattern, finishing touches, etc.) thanks to VR contents.

## II. Related Work

Passive haptics can be defined as the use of physical objects to provide feedback to users through their shape [25]. Several studies have shown that feeling the touch of physical objects in virtual environments can improve global immersion, knowledge about the spatial environment and users' sense of presence, particularly when these virtual objects react to touch just as their physical equivalents would [26]–[29].

To achieve satisfactory user experience in a virtual environment with passive haptics, the position of physical objects needs to be synchronized with virtual objects. It is also necessary to consider that perception of the size of a space and the position of the represented objects can be affected by several factors in a VR environment, such as technology or lack of an avatar, which may affect presence [30]–[33].

In a VRPH environment, users' haptic exploring can be done both passively and actively. With passive exploring, the surface reacts to touch and provides users with information. With active exploring, users explore the surface with their fingers and the palms of their hands. Recent studies have demonstrated that the second method facilitates users perceiving surfaces and helps them to better recognize the represented objects [34], [35]. This is why active exploring might be more suitable to evaluate consumer products.

Visual and haptic exploring strongly influences consumer product evaluations [36] and might also be relevant for online shopping experiences [37]. On the one hand, it has been demonstrated how a product's visual description can influence the opinion that consumers form about it and, thus, influences their purchasing decision [38], [39]. This visual information can also help consumers to mentally simulate how a product is used [38], [40] by, in turn, facilitating the appearance of product-related cognitive activities, which could impact product evaluations [41]. On the other hand, the haptic information that results from coming into physical contact with products can help consumers to form an opinion about them [42], [43], and can even improve consumers' capacity to evaluate their quality [44].

Although visual and haptic information can have a separate influence on how a product is perceived and evaluated, recent studies also demonstrate that some visual characteristics can influence how physical characteristics are perceived. Accordingly, research [45], [46] into how color (cold-warm) can be related to some physical properties, such as weight (light-heavy) or size (big-small), demonstrate that perceptual color experiences form part of the mental representation of tactile object attributes, and are applied to several fields like Tangible User Interfaces (TUIs).

To date, some works have investigated the different potentials of passive haptics. Lim and Follmer [47] created an application of small remote-control robots capable of transmitting physical sensations through several haptic patterns to different body parts, depending on the number of robots, movement or force of contact, among other parameters. Carvalheiro et al. [48] developed a sensors system to map users' hands and real objects, and to represent them in a synchronized manner in real time and in a virtual environment, which is useful for simulating physical interactions. Using low-resolution passive haptics combined with high-resolution VR images has enabled HMI dashboards to be developed [49] and to apply them to simulation booths in the aerospace sector [50], which can help to study how to reduce learning times. Other works have studied the importance of vibrotactile feedback on touchscreen devices [51]–[53] capable of returning confirmation feedback of a virtual button and transmitting meanings. Other research works have focused on physical objects capable of being reconfigured to adopt distinct basic physical shapes to be used as passive haptics in VR environments [54], [55].

Although some studies defend VR as a means to evaluate products in different development stages [56] [57], and others have analyzed the possibilities of distinct haptic devices to help evaluate products' usability via VR environments [58], very few works have either

studied the effect of haptic sensations on how a consumer evaluates a product presented by means of a VR environment or simultaneously compared this evaluation by other means. Our article attempts to extend knowledge in this field by comparing the evaluation of the same product by three different means: VR with visual, but no tactile inputs; VR with visual and tactile inputs (VRPH); the real product with visual and tactile inputs (R).

## III. Research Aim and Hypotheses

The present study aims to analyse the influence of different virtual presentation media (VR and VRPH) on how users perceive the product by comparing it to its traditional perception (a real product). To do so, a case study was done in which several users had to interact with a product in three settings. Evaluations of their perceived impressions in each setting were made using a semantic differential scale, which was subsequently analysed to detect any significant differences in users' evaluations.

This study posed the following hypotheses:

- H1: The medium used to present a piece of furniture influences how the users evaluate their perception of it.
- H2: The haptic interaction with the product (real or VRPH), as opposed to only the visual interaction (VR), influences the evaluation made of how users perceive the product.

## IV. Methods

### A. Case Study Approach

To test the posed hypotheses, a case study design was used in which the users had to interact with the same product, but it was presented by different media in such a way that each user could only interact with it by only one means.

The product selected to conduct the present study was a chair as it is a common piece of furniture with general characteristics known by all users. To enhance their haptic experience in some of design scenes, a round rug was placed below the chair so that when users moved closer to the product, they could stand on it and notice its touch.

With the means selected to present the product, the following scenes were created:

1. Scene Room 1 (SR1): Real environment, in which the product was placed along with some neutral physical furnishing elements to contextualise the scene. Users were able to see and touch the real product, but could neither touch any other element in the scene, nor move the product. They could stand on the real rug.
2. Scene Room 2 (SR2): VR simulated environment. A completely virtual setting represented by means of a VR headset. Users could see the product and the neutral furnishing elements by VR, could move around this scene, and even crouch to see hidden parts of the product, but could not touch anything. SR2 simulates all the SR1 conditions (furnishings, arrangement, lighting, etc), but via VR. In this scene, users could not stand on the real rug.
3. Scene Room 3 (SR3): VRPH simulated environment. A completely virtual-simulated setting represented via a VR headset, where the product to be studied was physically located. Users could see the product and the neutral furnishing elements via VR, move around the scene as in SR2, and touch any part of the product they had to evaluate without moving it. They could even sit on it, but could not touch any other element in the scene, except for the real rug. SR3 was exactly the same as SR2, but the product under study and the rug were physically added.

### B. Semantic Scale for Product Evaluations

To evaluate the product in each presentation setting, a semantic differential scale [59] was used based on bipolar pairs of adjectives about the product, which acted as product descriptors. Such scales are widely used to evaluate how products are perceived when many parameters need to be evaluated [60]–[62].

A semantic differential scale was created that contained 12 bipolar pairs of adjectives in Spanish, which was the mother tongue of the participants in the experimental phase. Researchers generally adapt a semantic differential scale in accordance with the nature of the product to be evaluated [63], and each researcher follows the research team's criterion to do so. As this criterion can be somewhat biased in some cases, the present study considered it more suitable for it to be based on a methodology already used by [64], [65]. This methodology sets three stages with which to draw up a list of bipolar pairs of adjectives by providing a list of the images of product examples taken from commercial websites (step 1) to then collect users' adjectives from these websites (step 2). Finally the adjectives are classified and filtered according to the four pleasure categories [12], [13] (step 3). Selecting the most common adjectives used to describe a chair according to Jordan's model allows us to take a representative sample of adjectives from each of the four categories (physio-pleasure, socio-pleasure, psycho-pleasure, ideo-pleasure), which provides us with information related to a wider spectrum of aspects that define the product, allowing us to aim for a more complete and global evaluation of the product. This may be of interest in order to better understand how the means of representation can influence some categories of adjectives more than others. On the contrary, if only the most common adjectives had been selected without considering these categories, the information obtained with the study could have been more limited in scope.

With this method, [64] drew up a semantic differential scale made up of five bipolar pairs of adjectives, and [65] prepared a scale made up of sixteen bipolar pairs. This methodology has been adapted to consider other variables to be able to obtain a suitable semantic differential scale with which to evaluate how an industrial product is perceived by the users or potential consumers of this product typology.

In our study, information was collected from four different sources (designers, users, manufacturers and distributors) so that the selection of bipolar pairs would match the more general criterion that adequately represents the descriptive terms employed by all the involved stakeholders. Eleven designers (9 men and 2 women, an average age of 35.8 years, with an average professional experience of 10 years) and 61 users (34 men and 27 women, with an average age of 21.8 years) were contacted and asked to answer a questionnaire. The websites of the manufacturers and distributors of the products in the studied category (12 in total: Ikea, Andreu World, Viccarbe Habitat, Cappellini, Cassina, Akaba, Barcelona Design, Gandia Blasco, De Padova, Bonaldo, Fornasarig, Amazon) were systematically analysed to collect the adjectives employed to describe the products.

To devise the questionnaire to be used by professional designers and users to collect the descriptive adjectives of the examples of products in the studied category, a search was done on websites specialising in the manufacturing or distribution of these products, which resulted in 50 images. Of these, the 15 most representative ones were selected from the whole studied product typologies range (Fig. 1). These images were edited to homogenise the way they appeared so that designers and users would not condition the way they looked. To collect adjectives, 15 examples were presented one by one using Google Forms, and five descriptive adjectives were requested of each presented example. Every participant was requested to make an evaluation about how much they liked each chair on a 5-interval Likert scale, where 1 was the lowest value ("I don't like it at all") and 5 was the highest value ("I like it very much").

It is worth pointing out that when completing questionnaires, designers and users did not need to make much effort with the first adjectives because they were generally the most evident characteristics of the presented product. In many cases however, the last two adjectives involved more effort as they were more singular and varied than the previous ones, and gave way to a richer more varied collection of terms. In this case, the collected adjectives had both positive ("nice", "elegant", etc.) and negative ("ugly", "uncomfortable", etc.) connotations. Likewise, it is worth stressing that no adjectives with negative connotations about products were given by manufacturers and distributors.


Fig. 1. Images presented in the questionnaires to describe the products of the studied category.

Having collected 5,611 adjectives (825 adjectives from 11 designers: each designer provided 75 adjectives, which were the result of writing 5 adjectives for each of the 15 chairs analysed; 4575 adjectives from 61 users; following the same procedure as the designers; 141 adjectives from 8 manufacturers' websites; 70 adjectives from 4 distributors' websites), the list was homogenised by eliminating their gender and number; that is, only the root of the term was considered, but differentiation in original sources was maintained. Then the frequency with which each adjective was repeated was counted, and those with the same meaning were grouped; e.g., "resistant" and "sturdy". Antonyms were also grouped to build the most frequent bipolar pairs of adjectives on the list by considering only the 25 most frequent ones from each source of origin. In those cases in which no antonym was available for one of the most frequent terms because they had only a positive or negative sense, they were added by the research team to create a bipolar pair, but no frequency value was added. To homogenise the order of magnitude of the frequency with which each source of origin was repeated (designers, users, etc.), the number of repetitions was weighted according to the sample of each source. Each resulting bipolar pair of adjectives was classified according to the four pleasures categories [12], [13] and placed in order of their frequency. The three most frequent ones from each category were selected. In order to ensure that when the semantic differential scale was used one of the extremes would not be taken as positive and the other as negative, some of the bipolar pairs of adjectives were randomly reversed. Finally a 7-interval scale was included, following a Likert scale, on all 12 bipolar pairs of adjectives (Table I) by taking 0 as a neutral value and 3 as the maximum value of both extremes. The purpose was to express that a higher value involved a greater extent of identifying the evaluated product with the corresponding adjective, but by avoiding taking one of the two extremes as being positive or negative. A consensus has been reached about this scale magnitude having a sufficient degree of reliability without users having difficulties to make evaluations [63].

TABLE I. List of the Bipolar Pairs of Selected Adjectives

| Physio | Psycho | Socio | Ideo |
|---|---|---|---|
| Comfortable - Uncomfortable | Practical - Useless | Modern - Classic | Elegant - Vulgar |
| Light - Heavy | Simple - Complex | Nice - Ugly | Handmade - Industrial |
| Resistant - Fragile | Versatile - Invariable | Overelaborate - Minimalist | Fun - Serious |

### C. Preparing Rooms for the Case Study

Three scenes were created in different rooms. In SR1, a series of real neutral furnishing elements was placed. The considered neutral furnishing elements came in basic forms, were white, grey or beige, and displayed no further decorative details. They included a medium-height shelving unit, two small pictures on the wall and a short-pile round rug placed beneath the product. The product selected for the case study was one of those selected to build the semantic differential scale (model 5 in Fig. 1). The beige-coloured Ikea Odger model was selected because, according to the evaluations made by designers and users when collecting adjectives to build the semantic differential scale, this model obtained a mean score compared to the rest of the sample.

With SR1, a 3D scene was modelled to generate SR2 and SR3. To model the virtual scene, the following tools were used: Solidworks 2018, with which building elements (walls, floor tiles, ceilings, lighting, etc.) and auxiliary furnishing elements (shelving unit, pictures and rug) were produced; Autodesk 3ds Max 2018, with which the product to be evaluated was generated and with which all the textures, colours, materials, lighting, etc., were included to make the scene as real as possible; Unity 2017.3.1f1, with which the executable VR model was generated to immerse users in the virtual room (SR2 and SR3).

Fig. 2 presents two images with which the similarity of SR1 (real) and SR2/SR3 (VR/VRPH) is shown. The high level of realism achieved in the virtual scene is stressed, which could have allowed the immersion sensation of all three scenes to be comparable.

Fig. 2. SR1 (real) on the left, SR2 (VR) / SR3 (VRPH) on the right.

The equipment employed in SR2 and SR3 consisted in a graphics workstation (HP Z420 Workstation x64, Intel Xeon processor CPU E5-1660 v2 @ 3.70GHz, 6-CPU Core, 32GB RAM and NVIDIA Quadro K5000 graphics card), a Oculus Rift VR headset, two position sensors placed at the front of the scene and two Oculus Touch controllers, which were employed to only calibrate the scene.

### D. Sample (Participants)

To run the experimental phase, 203 voluntary users participated. Gender distribution was 111 men and 92 women aged between 18 and 40 years, with a mean age of 22.77 years. All the voluntary participants were studying the Degree in Industrial Design and Product Development Engineering. Regarding sample size calculation, a priori power analysis was conducted with G*Power [66] supposing

an one-way ANOVA statistical test with these input parameters: effect size: 0.25, α=0.05, (1-β)=0.85 and 3 groups. G*Power provided a total sample size of 180. In order to guarantee to achieve at least a power of 0.85 as used with G*Power, the total sample size used in the experiment was 203. Although finally a Kruskal Wallis test has been applied due to the data non-normality, we are confident that a power of 0.85 is achieved considering that with non-symmetrical distributions the non-parametrical Kruskal-Wallis test results in a higher power compared to the classical one-way ANOVA [67].

An initial survey was conducted with the participants to learn about their experience with VR devices: 96 (47.29%) users had no experience, 98 (48.28%) had some former experience and only 7 (3.44%) stated they were very familiar with VR devices. Two people did not answer this survey question (0.99%).

### E. The Experiment Protocol

The experiment was carried out on 3 days of one same week to limit as much as possible comments about the actions performed in the experiment among users who might know one another. Participants were also asked to maintain the confidentiality of the actions they performed, at least until the experimental phase has ended.

To design the experimental phase, the sample was divided into three groups of users according to the built Scene rooms: Group 1, R (65); Group 2, VR (68); Group 3, VRPH (70).

To build the Scene rooms, two rooms were used whose size and characteristics were similar. Each scene was configured according to the presented conditions. The same room was used for SR2 and SR3, with the only difference appearing in SR3 (VRPH), with a rug and a real chair standing in the centre so that users could touch the chair. In SR2 (VR) all the elements were virtual and, hence, the real chair and rug were removed.

A protocol was written to perform the experiment in all the Scene rooms so that the sequence of steps to follow or the indications students had to do were independent of the researcher involved in each case.

The experiment's sequence was as follows:

#### 1. Stage 1. Welcome Room (2 Min.)

In order to preserve the figures' integrity across multiple computer platforms, we accept files in the following formats: .EPS/.PDF/.PS/.AI. All fonts must be embedded or text converted to outlines in order to achieve the best-quality results.

Step 1. The users came to the Welcome room (outside the Scene rooms), were identified to determine the as-signed Scene room and signed an informed consent to participate in the experiment. An informal friendly conversation was held so that the participants would not feel worried and they were accompanied to the corresponding Scene room.

#### 2. Stage 2. Scene Rooms (5 Min.)

Step 2. In SR2 and SR3, a VR headset was placed and adjusted to each user's characteristics. Under no circumstances were users allowed to previously view the real scene at any time to avoid conditioning their subsequent evaluation. To do so, a screen was positioned to separate the area in which the VR headset was placed from the scene. The participants were explained that they were about to see a VR scene, they had to respect the room's limits and a researcher would be at their side at all times to avoid them becoming entangled with the VR headset cable. In SR1, they simply entered the room.

Step 3. All the users were explained that they had to observe (and touch or sit on in SR1 and SR3) the product (the chair) located in the scene; they could move around it, crouch or move closer to see details. They were also explained that when the observation phase had ended, they would be handed a survey to give their opinion about the product's characteristics.

Step 4. Each user had 2 minutes to experiment with the product in accordance with the conditions of each scene.

Step 5. In SR2 and SR3, the VR headset was removed behind the screen. The participants were asked about their first impression or any outstanding observation, and they were asked to leave the survey room to complete the survey and, if necessary, to provide details about the aforementioned observations.

#### 3. Stage 3. Survey Room (5 Min.)

Step 6. All the users completed the survey without saying anything to anyone. A researcher remained to explain any doubt they had about the survey's questions.

In the questionnaire employed to evaluate the studied product in each Scene some questions were included about possible viewing problems that users may have had (myopia, astigmatism, use of glasses or contact lenses), which could have conditioned the use of the VR headset according to previous experience with VR technology. The participants had to rate the chair in accordance with all 12 semantic pairs using a 7-point semantic differential scale ("Rate the chair you just saw according to whether you think it is closer or further away from the following adjectives"). Next they had to indicate how much they liked the chair globally by scoring their answer on a 5-point Likert scale that went from 1 ("I do not like it at all") to 5 ("I like it very much"). An open space was left for the participants to include comments about the experience.

Fig. 3 provides some examples of users in step 4 of stage 2. The layout of the equipment utilised for the experiment in both scene rooms VR and VRPH is seen, namely the position sensors of the VR headset. The cable linking the VR headset and the PC is shown, which the researchers had to supervise at all times so that the users were neither entangled nor damaged equipment.

In both SR1 and SR3, the users could touch the product they had to evaluate, and they even sat on it, but were asked to not move it from where it was placed.

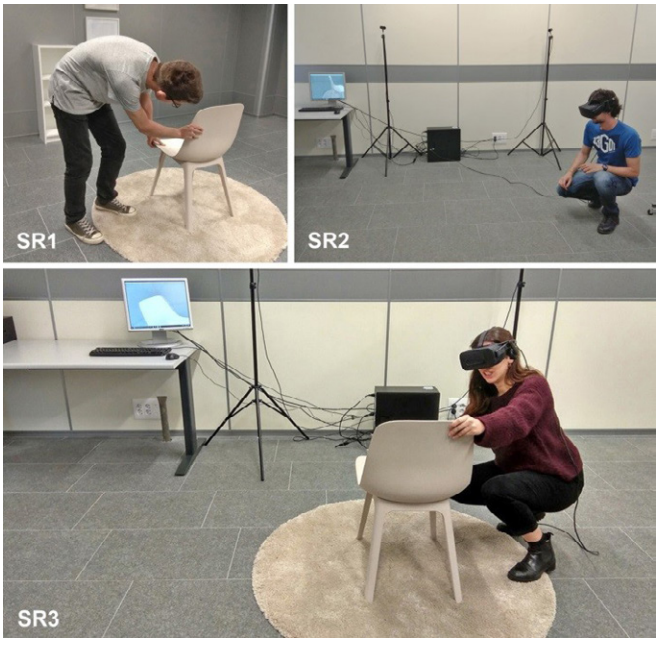



Fig. 3. Participants during the experiment in the different scene rooms (SR1: R, SR2: VR, SR3: VRPH).

## V. Results

To help interpret the data collected by surveys in each setting (R, VR and VRPH), an inferential statistical method was used with which the posed hypotheses were tested. It was possible to distinguish two collected datasets: those corresponding to the differential semantic scale evaluations for each semantic pair of adjectives, and those corresponding to the "I like it" evaluation.

Regarding the first dataset, Table II includes the descriptive statistics for differential semantic scales. It is noteworthy that data were collected by a 7-interval Likert scale with a central neutral value of 0 and two extreme values of 3 (in absolute values), where a higher value indicates a better correspondence with the adjective represented on this extreme. For suitable data processing, a negative value to the left of the survey was taken to simply indicate that the adjective on this extreme came closer and had no further connotation. As the scale values were discrete, the value of the median was also discrete. Thus the values of the means and standard deviations are also indicated because they may better represent the distribution of the collected value.

TABLE II. Descriptive Statistics for Differential Semantic Scales

| Semantic scales | | Conditions R | VR | VRPH |
|---|---|---|---|---|
| Uncomfortable - **Comfortable** | Mean | 1.09 | *.63* | **1.87** |
| | Median | 2.00 | 1.00 | 2.00 |
| | Std. Deviation | 1.51 | 1.23 | 1.15 |
| **Light** - Heavy | Mean | *-.75* | **-1.32** | -1.09 |
| | Median | -1.00 | -2.00 | -1.00 |
| | Std. Deviation | 1.40 | 1.25 | 1.35 |
| Fragile - **Resistant** | Mean | 1.05 | *.62* | **1.66** |
| | Median | 1.00 | 1.00 | 2.00 |
| | Std. Deviation | 1.23 | 1.28 | 1.31 |
| **Practical** - Useless | Mean | -1.55 | *-1.41* | **-1.79** |
| | Median | -2.00 | -2.00 | -2.00 |
| | Std. Deviation | 1.13 | 1.12 | 1.27 |
| Complex - **Simple** | Mean | *1.68* | 1.75 | **1.87** |
| | Median | 2.00 | 2.00 | 2.00 |
| | Std. Deviation | 1.45 | 1.27 | 1.55 |
| **Invariable - Versatile** | Mean | **.22** | **.16** | **.23** |
| | Median | .00 | .00 | .00 |
| | Std. Deviation | 1.64 | 1.47 | 1.79 |
| **Modern** - Classic | Mean | -1.06 | *-.99* | **-1.27** |
| | Median | -1.00 | -1.00 | -2.00 |
| | Std. Deviation | 1.33 | 1.35 | 1.37 |
| Ugly - **Nice** | Mean | *.80* | 1.25 | **1.66** |
| | Median | 1.00 | 2.00 | 2.00 |
| | Std. Deviation | 1.52 | 1.43 | 1.34 |
| **Minimalist** - Overelaborate | Mean | -1.75 | *-1.53* | **-1.91** |
| | Median | -2.00 | -2.00 | -2.00 |
| | Std. Deviation | 1.00 | 1.23 | 1.06 |
| Vulgar - **Elegant** | Mean | *.71* | .97 | **1.56** |
| | Median | 1.00 | 1.00 | 2.00 |
| | Std. Deviation | 1.31 | 1.17 | 1.23 |
| Handmade - **Industrial** | Mean | 1.60 | *1.06* | **1.84** |
| | Median | 2.00 | 1.00 | 2.00 |
| | Std. Deviation | 1.36 | 1.51 | 1.29 |
| **Serious - Fun** | Mean | **-.35** | **.32** | **.27** |
| | Median | .00 | .00 | .00 |
| | Std. Deviation | 1.24 | 1.29 | 1.38 |

Highest values and corresponding adjective in bold, lowest values in italics.

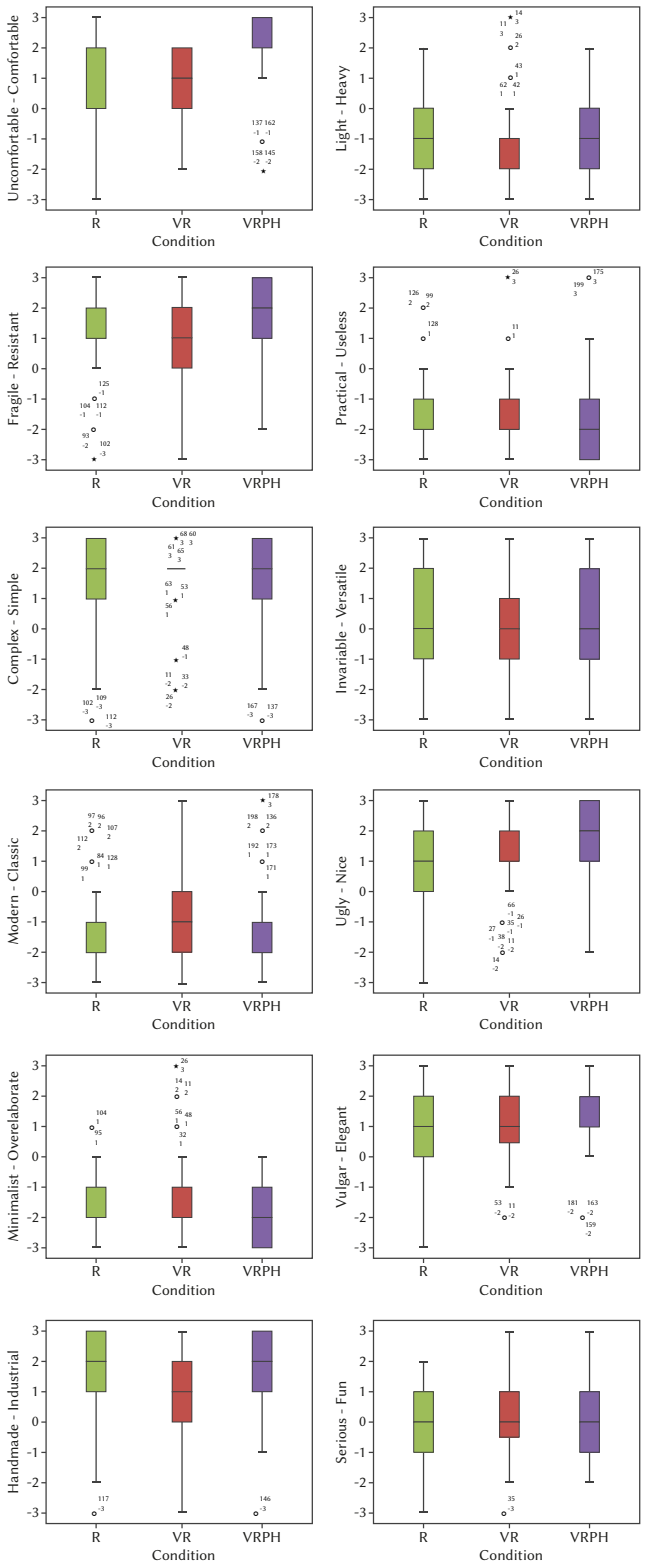



Fig. 4. Box plots for the semantic scales.

There were two semantic differentials for which users gave similar scores for all three conditions, which came very close to the neutral score (0), namely semantic differentials "versatile-invariable" and "fun-serious". For the remarks collected by the evaluators in stage 3 in the survey room, it was the adjectives that made users doubt the most when relating them to the presented product, which could have led to a poorly polarized neutral score.

Therefore, by contemplating only the study of the means and standard deviations, we could consider that the score for the product presented in VRPH was more positive than those given in VR and R.

Fig. 4 presents the box plots for the semantic scales, showing the distribution of the values of the collected samples for all the semantic pairs for all the studied conditions. It is noteworthy that as the discrete values corresponded to a reduced 7-interval scale, both the box plots and whiskers took the positions of the integers corresponding to this interval. Thus their interpretation had to be done by complementing with other values such as mean or standard deviation.

TABLE III. Descriptive Statistics for the Overall Evaluation Scales

| Overall Evaluation | | Conditions R | VR | VRPH |
|---|---|---|---|---|
| I like it (1-5) | Mean | **3.95** | *3.84* | 3.86 |
| | Median | 4.00 | 4.00 | 4.00 |
| | Std. Deviation | .76 | .56 | .62 |

Highest values in bold, lowest values in italics.

TABLE IV. Test of Normality of the Differential Semantic Scale

| Semantic scales | Cond. | Kolmogorov-Smirnov[a] Stat. | df | Sig. | Shapiro-Wilk Stat. | df | Sig. |
|---|---|---|---|---|---|---|---|
| Uncomfortable - Comfortable | R | .250 | 65 | .000 | .869 | 65 | .000 |
| | VRPH | .344 | 70 | .000 | .714 | 70 | .000 |
| | VR | .220 | 68 | .000 | .867 | 68 | .000 |
| Light - Heavy | R | .185 | 65 | .000 | .908 | 65 | .000 |
| | VRPH | .223 | 70 | .000 | .901 | 70 | .000 |
| | VR | .294 | 68 | .000 | .754 | 68 | .000 |
| Fragile - Resistant | R | .239 | 65 | .000 | .896 | 65 | .000 |
| | VRPH | .289 | 70 | .000 | .814 | 70 | .000 |
| | VR | .156 | 68 | .000 | .941 | 68 | .003 |
| Practical - Useless | R | .238 | 65 | .000 | .865 | 65 | .000 |
| | VRPH | .295 | 70 | .000 | .774 | 70 | .000 |
| | VR | .215 | 68 | .000 | .879 | 68 | .000 |
| Complex - Simple | R | .311 | 65 | .000 | .733 | 65 | .000 |
| | VRPH | .262 | 70 | .000 | .721 | 70 | .000 |
| | VR | .357 | 68 | .000 | .694 | 68 | .000 |
| Invariable - Versatile | R | .156 | 65 | .000 | .939 | 65 | .003 |
| | VRPH | .152 | 70 | .000 | .928 | 70 | .001 |
| | VR | .162 | 68 | .000 | .931 | 68 | .001 |
| Modern - Classic | R | .251 | 65 | .000 | .886 | 65 | .000 |
| | VRPH | .222 | 70 | .000 | .875 | 70 | .000 |
| | VR | .240 | 68 | .000 | .883 | 68 | .000 |
| Ugly - Nice | R | .229 | 65 | .000 | .894 | 65 | .000 |
| | VRPH | .315 | 70 | .000 | .798 | 70 | .000 |
| | VR | .274 | 68 | .000 | .832 | 68 | .000 |
| Minimalist - Overelaborate | R | .289 | 65 | .000 | .856 | 65 | .000 |
| | VRPH | .219 | 70 | .000 | .830 | 70 | .000 |
| | VR | .311 | 68 | .000 | .771 | 68 | .000 |
| Vulgar - Elegant | R | .188 | 65 | .000 | .926 | 65 | .001 |
| | VRPH | .326 | 70 | .000 | .753 | 70 | .000 |
| | VR | .260 | 68 | .000 | .897 | 68 | .000 |
| Handmade - Industrial | R | .216 | 65 | .000 | .863 | 65 | .000 |
| | VRPH | .248 | 70 | .000 | .817 | 70 | .000 |
| | VR | .189 | 68 | .000 | .906 | 68 | .000 |
| Serious - Fun | R | .181 | 65 | .000 | .922 | 65 | .001 |
| | VRPH | .144 | 70 | .001 | .935 | 70 | .001 |
| | VR | .171 | 68 | .000 | .944 | 68 | .004 |

[a] Lilliefors Significance Correction.

TABLE V. Test of Normality of the "I Like It" Question

| | Cond. | Kolmogorov-Smirnov[a] Stat. | df | Sig. | Shapiro-Wilk Stat. | df | Sig. |
|---|---|---|---|---|---|---|---|
| I like it (1-5) | R | .324 | 65 | .000 | .791 | 65 | .000 |
| | VRPH | .348 | 70 | .000 | .778 | 70 | .000 |
| | VR | .393 | 68 | .000 | .721 | 68 | .000 |

[a] Lilliefors Significance Correction.

TABLE VI. Ranks and Kruskal-Wallis Test

| | Semantic scales | Cond. | N | Mean Rank | Kruskal-Wallis Test |
|---|---|---|---|---|---|
| PHYSIO | **Uncomfortable - Comfortable** | R | 65 | 98.50 | |
| | | VRPH | 70 | 132.41 | **$X^2(2)=37.627$** |
| | | VR | 68 | 74.04 | **p<.001** |
| | | Total | 203 | | |
| | **Light - Heavy** | R | 65 | 115.49 | |
| | | VRPH | 70 | 101.41 | **X2(2)=6.981** |
| | | VR | 68 | 89.71 | **p=.030** |
| | | Total | 203 | | |
| | **Fragile - Resistant** | R | 65 | 97.71 | |
| | | VRPH | 70 | 128.55 | **$X^2(2)=26.801$** |
| | | VR | 68 | 78.77 | **p<.001** |
| | | Total | 203 | | |
| PSYCHO | **Practical - Useless** | R | 65 | 104.78 | |
| | | VRPH | 70 | 88.44 | **$X^2(2)=7.029$** |
| | | VR | 68 | 113.31 | **p=.030** |
| | | Total | 203 | | |
| | Complex - Simple | R | 65 | 96.56 | |
| | | VRPH | 70 | 111.53 | $X^2(2)=3.179$ |
| | | VR | 68 | 97.39 | $p=.204$ |
| | | Total | 203 | | |
| | Invariable - Versatile | R | 65 | 102.30 | |
| | | VRPH | 70 | 103.57 | $X^2(2)=.128$ |
| | | VR | 68 | 100.10 | $p=.938$ |
| | | Total | 203 | | |
| SOCIO | Modern - Classic | R | 65 | 106.25 | |
| | | VRPH | 70 | 93.57 | $X^2(2)=2.368$ |
| | | VR | 68 | 106.61 | $p=.306$ |
| | | Total | 203 | | |
| | **Ugly - Nice** | R | 65 | 82.52 | |
| | | VRPH | 70 | 120.46 | **$X^2(2)=15.278$** |
| | | VR | 68 | 101.62 | **p<.001** |
| | | Total | 203 | | |
| | Minimalist - Overelaborate | R | 65 | 102.69 | |
| | | VRPH | 70 | 92.56 | $X^2(2)=3.842$ |
| | | VR | 68 | 111.06 | $p=.146$ |
| | | Total | 203 | | |
| IDEO | **Vulgar - Elegant** | R | 65 | 83.05 | |
| | | VRPH | 70 | 127.79 | **$X^2(2)=23.364$** |
| | | VR | 68 | 93.57 | **p<.001** |
| | | Total | 203 | | |
| | **Handmade - Industrial** | R | 65 | 105.29 | |
| | | VRPH | 70 | 116.55 | **$X^2(2)=11.672$** |
| | | VR | 68 | 83.88 | **p=.003** |
| | | Total | 203 | | |
| | **Serious - Fun** | R | 65 | 84.12 | |
| | | VRPH | 70 | 108.70 | **$X^2(2)=9.455$** |
| | | VR | 68 | 112.20 | **p=.009** |
| | | Total | 203 | | |

Semantic scales in bold to identify the scales with statistically significant differences.

Kolmogorov-Smirnov and Shapiro-Wilk tests (α=0.05) (Table IV) showed that semantic scales did not follow a normal distribution in all cases. Consequently, an ANOVA test proved unsuitable for testing, and Kruskal-Wallis was selected. This is a non-parametric method for testing whether samples originate from the same distribution. The viewing conditions were taken as the independent variables (R, VR and VRPH) and the scores for each semantic pair as the dependent variables.

The null hypothesis of the Kruskal-Wallis test stated that the mean ranks of semantic scales scores in the three experimental conditions were the same. Firstly, four assumptions had to be checked:

1. The dependent variable should be measured at the ordinal or continuous level. In our case, semantic scale scores were measured from -3 to 3.
2. The independent variable should consist of two categorical independent groups or more. In our case, we had three independent groups (R, VR and VRPH).
3. There was no relationship between the observations in each group or between the groups themselves.
4. The distributions in each group should have a similar shape and variability (as seen in Fig. 4).

The Kruskal-Wallis test results (Table VI) revealed that the null hypothesis was not confirmed (significance level .05) on many semantic scales. As shown in bold, significant differences appeared among some semantic differentials when comparing the three conditions.

In order to study if these significant differences appeared among the three conditions or were due to differences between them, Dunn's post hoc test was performed (a=0.05) using the adjustment p-value to make a pairwise comparison according to the Kruskal-Wallis test. The results are shown in Table VII. As multiple tests were carried out, Bonferroni adjustment was applied to all the Dunn's p-values, as presented in the last column of Table VII (adjusted p-value), which also includes the effect size calculated as:

$r = z/\sqrt{N}$ N = total number of observations.

In Table VII, there are 13 pairwise comparisons for which there are statistically significant differences, and the VRPH condition is present in 10 of these pairs. Only "Uncomfortable-Comfortable" presented significant differences for all three use conditions grouped into pairs. The VRPH score was better than the other two, while the R score was better than that for VR.

For "Light-Heavy", differences were found only between R and VR, where VR was better. Thus no differences appeared between users' scores for "Light Heavy" when comparing VRPH with the other two conditions, so the product's evaluation was not harmed. Conversely, differences were observed between users' evaluations for "Fragile-Resistant" or "Vulgar-Elegant" when comparing VRPH to the other two conditions as their score for VRPH was better, as previously observed (Table II and Fig. 4). Table VII also shows that significant differences were found only when comparing VRPH to either of the other two conditions, such as "Practical Useless", "Ugly-Nice" or "Handmade-Industrial", with the best score going to VRPH.

Regarding the second dataset, corresponding to the "I like it" question, Table III includes the descriptive statistics for the overall evaluation as regards the question "I like it". In this case, the employed evaluation scale was a 5-interval Likert scale, use minimum value was 1 and its maximum value was 5.

The statistical descriptives in Table III reveal that on a scale from 1 to 5, the scores of all three conditions come very close and are slightly higher in R and lower in VR. In Fig. 5, the box plot figure only identifies the medians and lots of outliers because more than 50% of the distribution of values takes a value of 4. Thus it was not possible to draw boxes in the figure.

To study the differences between the scores for the question "I like it", due to the data non-normality (Table V), a Kruskal-Wallis test was applied. It showed that there was no statistically significant difference between the three viewing conditions (R, VR, VRPH), $X^2(2)=2.085$, p=.353.

TABLE VII. Dunn's Post Hoc Tests

| | Semantic Scale | Pairwise comparison | Diff. between mean ranks | Std. Error | z | r | p | Adjust.p |
|---|---|---|---|---|---|---|---|---|
| Physio | Uncomfort.-Comfortable | **R-VRPH** | -33.907 | 9.673 | -3.505 | 0.302 | .000 | .001 |
| | | **VR-R** | 24.456 | 9.741 | 2.511 | 0.218 | .012 | .036 |
| | | **VR- VRPH** | 58.363 | 9.562 | 6.104 | 0.520 | .000 | .000 |
| | Light -Heavy | R-VRPH | -14.085 | 9.697 | -1.453 | 0.125 | .146 | .439 |
| | | **VR-R** | 25.779 | 9.765 | 2.640 | 0.225 | .008 | .025 |
| | | VR- VRPH | 11.694 | 9.585 | 1.220 | 0.106 | .222 | .667 |
| | Fragile - Resistant | **R- VRPH** | -30.842 | 9.827 | -3.139 | 0.270 | .002 | .005 |
| | | VR-R | 18.936 | 9.896 | 1.913 | 0.166 | .056 | .167 |
| | | **VR- VRPH** | 49.778 | 9.714 | 5.124 | 0.436 | .000 | .000 |
| Psycho | Practical - Useless | R- VRPH | -16.341 | 9.653 | -1.693 | 0.147 | .090 | .271 |
| | | VR-R | 8.532 | 9.721 | .878 | 0.076 | .380 | 1.000 |
| | | **VR- VRPH** | 24.873 | 9.542 | 2.607 | 0.222 | .009 | .027 |
| Socio | Ugly - Nice | **R- VRPH** | -37.934 | 9.707 | -3.908 | 0.333 | .000 | .000 |
| | | VR-R | 19.095 | 9.775 | 1.953 | 0.169 | .051 | .152 |
| | | VR- VRPH | 18.839 | 9.595 | 1.964 | 0.169 | .050 | .149 |
| Ideo | Vulgar - Elegant | **R- VRPH** | -44.740 | 9.741 | -4.593 | 0.391 | .000 | .000 |
| | | VR-R | 10.527 | 9.810 | 1.073 | 0.093 | .283 | .850 |
| | | **VR- VRPH** | 34.212 | 9.629 | 3.553 | 0.306 | .000 | .001 |
| | Handmade - Industrial | R- VRPH | -11.258 | 9.811 | -1.148 | 0.099 | .251 | .754 |
| | | VR-R | 21.417 | 9.880 | 2.168 | 0.188 | .030 | .091 |
| | | **VR- VRPH** | 32.675 | 9.698 | 3.369 | 0.287 | .001 | .002 |
| | Serious - Fun | **R- VRPH** | -24.585 | 9.864 | -2.492 | 0.216 | .013 | .038 |
| | | **VR-R** | 28.083 | 9.934 | 2.827 | 0.241 | .005 | .014 |
| | | VR- VRPH | 3.499 | 9.751 | .359 | 0.031 | .720 | 1.000 |

Pairs in bold identify significant differences when applying a .05 significant level using the adjusted p-value.

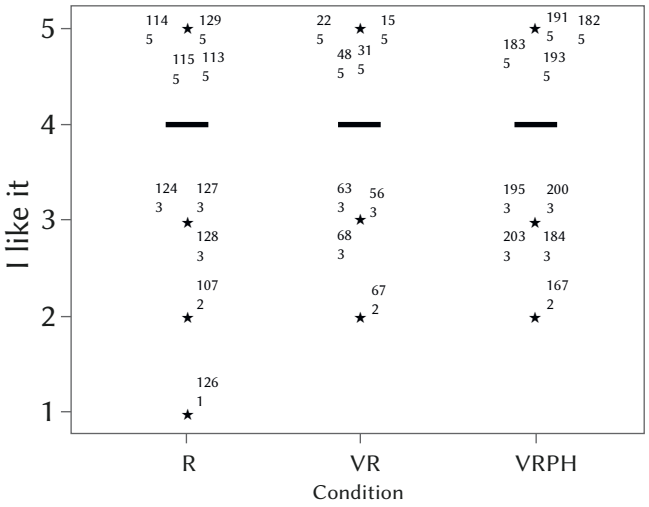



Fig. 5. Box plot for the overall evaluation.

## VI. Discussion

As Fig. 5 reveals, the overall evaluation that users made of the presented furnishing product is practically the same. So we conclude that the means employed to present this particular product does not influence its overall evaluation.

When the evaluations of the semantic differentials were analyzed for the three studied conditions (Table II), VRPH was highlighted in most cases with a higher mean, as former works have found using a similar product [60]. In Table VII we see that the biggest significant differences in the pairwise comparison tended to appear between VRPH and one of the other two means, and VRPH was evaluated the best. So we conclude that when this furnishing product is presented by VRPH, users evaluate it more favorably than for the R or VR condition. As the statistical differences refer only to the tested product, we cannot confidently transfer these results to other chair models or product categories. But if the results were similar in other further studies evaluating other products, this could mean advantages when presenting a product in a showroom because potential buyers could gain a better impression of it. Yet if we were to employ this means in the design phase to predict future users' responses, mistaken design decisions might be made and might lead to a product being developed that could be evaluated worse by users when presented in other means.

According to the results in Table VI, some semantic differentials present significant differences depending on the viewing conditions. Thus using one means or another to present this product will depend on the category in which evaluations with reliable results we wish to obtain.

The semantic differentials that correspond to categories Physio and Ideo are those with the most significant differences on the whole when comparing the various viewing conditions, where VRPH is the condition that obtained the best valuations. If we bear in mind that Physio refers to the pleasures deriving from sensorial organs like touch, and Ideo refers to esthetic values, it would be logical to think that resorting to passive haptics in an interaction condition to evaluate furnishing products would lead to a better evaluation of the related semantic differentials. So VRPH would be more suitable for presenting furnishing products, where the tactile interaction and the esthetic value are important. However, since this study is limited to a single model of chair, it would be necessary to carry out other studies to check that these conclusions are also applicable to other furniture products.

On the other hand, in the categories of Socio, with a social connotation, and Ideo, with an emotional connotation, no viewing condition would stand out from the rest. Therefore, these results may suggest that VR could be used to present products with a social or emotional character without harming users' evaluations. So making a physical product available for it to be evaluated would not be necessary as this could be done from home without going a physical store. This could also be useful in some design process phases to evaluate product alternatives without having to build a physical prototype. However, given the limitations of this study, these assumptions need to be tested by further research.

## VII. Conclusions

What the present study demonstrates is that the ways by which this piece of furniture is presented (R, RV, VRPH) influences how users perceive it (H1). It also demonstrates that differences are found between the score of perceiving this product presented in a virtual means (VR or VRPH) and presented in a physical means (R). Finally, it demonstrates that users' haptic interaction with this product (R or VRPH), as opposed to only their visual interaction (VR), influences how users perceive and evaluate it (H2).

The semantic pairs selection was done along with grouping them into four categories according to Jordan's model to run an accurate analysis of how presentation means influences users' responses. It is worth stressing that the experimental results showed that not all four categories performed the same with variation in presentation type. This is very important if we wish to use these technologies in the initial design cycle phase where the purposes are to evaluate several design alternatives, and to use VR technologies to avoid building physical prototypes and to speed up decision making. However, the experiment would need to be extended to include other product samples and categories before these conclusions could be generalised.

The scores made when observing this product in VRPH were higher, as evidenced by the R and VR scores in most semantic pairs. So it is worth stressing that observing and interacting with products using VRPH could result in the product being positively valued, which could favor purchasing decisions. VRPH was also the means in which certain physical characteristics of this chair, like "comfortable" or "resistant" (Physio), were more positively evaluated than in a means in which touching is not allowed (VR). Thus using VRPH as a presentation means seems suitable if higher evaluations are sought of consumer products that relate well to physical and tactile characteristics, like chairs, but it does not necessarily have to more positively influence the evaluation of products with marked social (Socio) or emotional (Ideo) characteristics than other means.

Again, the conclusions drawn from this study are limited by the fact that only one product was analysed. Further research is therefore necessary to draw more general conclusions that can also be applied to other similar products, or to other product categories.

## Acknowledgment

This work was supported by the Spanish Ministry of Science and Innovation (grant number PID2019-106426RB-C32) and the Universitat Jaume I (grant number UJI-B2019-39). The authors also wish to thank the designers Francisco Cueto, Paula Bañuelos and Cristina Gasch for their collaboration.

## References

[1] Z. Jiang and I. Benbasat, "Investigating the influence of the functional mechanisms of online product presentations," *Information Systems Research*, vol. 18, no. 4, pp. 454-470, 2007, doi:10.1287/isre.1070.0124.

[2] S.W. Jeong, A.M. Fiore, L.S. Niehm and F.O. Lorenz, "The role of experiential value in online shopping: The impacts of product presentation on consumer responses towards an apparel web site," *Internet Research*, vol. 19, no. 1, pp. 105-124, 2009, doi:10.1108/10662240910927858.

[3] J. Yoo, and M. Kim, "The effects of online product presentation on consumer responses: A mental imagery perspective," *Journal of Business Research*, vol. 67, no. 11, pp. 2464-2472, 2014, doi:10.1016/j.jbusres.2014.03.006.

[4] T.G. Saraswati, "Driving Factors of Consumer to Purchase Furniture Online on IKEA Indonesia Website," *Jurnal Sekretaris & Administrasi Bisnis*, vol. 2, no. 1, pp. 19-28, 2018.

[5] S. Unal, "The Mediating Role of Product Familiarity in Consumer Animosity", *Journal of Accounting and Marketing*, vol. 6, no. 4, pp. 1-11, 2017, doi:10.4172/2168-9601.1000257.

[6] N.A.A. Jalil, A. Fikry and A. Zainuddin, "The Impact of Store Atmospherics, Perceived Value, and Customer Satisfaction on Behavioural Intention", Procedia Economics and Finance, vol. 37, pp. 538-544, 2016, doi:10.1016/S2212-5671(16)30162-9.

[7] E. Naderi, I. Naderi and B. Balakrishnan, "Product design matters, but is it enough? Consumers' responses to product design and environment congruence," *Journal of Product & Brand Management*, vol. 29, no. 7, pp. 939-954, doi: 10.1108/JPBM-08-2018-1975.

[8] Y. Chen, "Neurological Effect of the Aesthetics of Product Design on the Decision-making Process of Consumers," *NeuroQuantology*, vol. 16, no. 6, 2018.

[9] M.B. Holbrook and E.C. Hirschman, "The experiential aspects of consumption: Consumer fantasies, feelings, and fun," *Journal of consumer research*, vol. 9, no. 2, pp. 132-140, 1982.

[10] T. H. Dodd and A.W. Gustafson, "Product, environmental, and service attributes that influence consumer attitudes and purchases at wineries," *Journal of Food Products Marketing*, vol. 4, no. 3, pp. 41-59, 1997.

[11] R.E. Petty, R.H. Unnava and A.J. Strathman (1991). "Theories of attitude change," in *Handbook of Consumer Behavior*, T.S. Robertson and H.H. Kassarjian, eds., Englewood Cliffs, NJ: Prentice-Hall, pp. 241-280, 1991.

[12] P. Jordan, *Designing Pleasurable Products: An Introduction to the New Human Factors*, London: Taylor & Francis, 2000.

[13] L. Tiger, *The Pursuit of Pleasure*, Boston: Little, Brown & Company, 1992.

[14] K. Wu, J. Vassileva, Y. Zhao, Z. Noorian, W. Waldner and I. Adaji, "Complexity or simplicity? Designing product pictures for advertising in online marketplaces," *Journal of Retailing and Consumer Services*, vol. 28, pp. 17-27, 2016, doi:10.1016/j.jretconser.2015.08.009.

[15] 15 C. Flavián, R. Gurrea and C. Orús, "The Impact of Online Product Presentation on Consumers' Perceptions: An Experimental Analysis," *International Journal of E-Services and Mobile Applications*, vol. 1, no. 3, pp. 17-37, 2009, doi:10.4018/jesma.2009070102.

[16] M.A. Artacho-Ramírez MA, J.A. Diego-Mas JA and J. Alcaide-Marzal, "Influence of the mode of graphical representation on the perception of product aesthetic and emotional features: an exploratory study," *Int. J. Ind. Ergon.* vol. 38, pp. 942–952, 2008, doi:10.1016/j.ergon .2008.02.020.

[17] E. Naderi, I. Naderi, I. and B. Balakrishnan, "Product design matters, but is it enough? Consumers' responses to product design and environment congruence," *Journal of Product & Brand Management*, 2020, doi:10.1108/JPBM-08-2018-1975.

[18] T. Verhagen, C. Vonkeman, F. Feldberg, and P. Verhagen, "Present it like it is here: Creating local presence to improve online product experiences," *Computers in human behavior*, vol. 39, pp. 270-280, 2014.

[19] K.S. Suh and Y.E. Lee, "The effects of virtual reality on consumer learning: an empirical investigation," *Management Information Systems Quarterly*, vol. 29, pp. 673-697, 2005, doi:10.2307/25148705.

[20] D. Grewal, S. M. Noble, A.L. Roggeveen and J. Nordfalt, "The future of in-store technology," *Journal of the Academy of Marketing Science*, vol. 48, no. 1, pp. 96-113, 2020, doi:10.1007/s11747-019-00697-z.

[21] S. Steinmann, T. Kilian and D. Brylla, "Experiencing Products Virtually: The Role of Vividness and Interactivity in Influencing Mental Imagery and User Reactions," *Proceedings of the 35th International Conference on Information Systems (ICIS)*, pp. 1-20, 2014.

[22] C.-J. Keng, T.H. Liao, and Y.I. Yang, "The effects of sequential combinations of virtual experience, direct experience, and indirect experience: the moderating roles of need for touch and product involvement," *Electronic Commerce Research*, vol. 12, no. 2, pp. 177-199, 2012, doi:10.1007/s10660-012-9093-9.

[23] J. Peck and J. Wiggins, "It just feels good: Customers' affective response to touch and its influence on persuasion," *Journal of Marketing*, vol. 70, no. 4, 2006, pp. 56-69.

[24] A. Zenner, F. Kosmalla, J. Ehrlich, P. Hell, G. Kahl, C. Murlowski, M. Speicher, F. Daiber, D. Heinrich and A. Krüger, "A Virtual Reality Couch Configurator Leveraging Passive Haptic Feedback," *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*, pp. 1-8, 2020, doi: 10.1145/3334480.3382953.

[25] R.W. Lindeman, J.L. Sibert and J.K. Hahn, "Hand-Held Windows: Towards Effective 2D Interaction in Immersive Virtual Environments," *Proceedings of the IEEE Virtual Reality (VR '99)*, pp. 205-212, 1999, doi: 10.1109/VR.1999.756952.

[26] B.E. Insko, "Passive haptics significantly enhances virtual environments," PhD dissertation, The University of North North Carolina, 2001.

[27] N. Tardif, C.E. Therrien and S. Bouchard, "Re-Examining Psychological Mechanisms Underlying Virtual Reality-Based Exposure for Spider Phobia," *Cyberpsychology, Behavior and Social Networking*, vol. 22, no. 1, pp. 39-45, 2019, doi: 10.1089/cyber.2017.0711.

[28] A. S. Carlin, H. G. Hoffman and S. Weghorst, "Virtual reality and tactile augmentation in the treatment of spider phobia: A case report," *Behaviour Res. Therapy*, vol. 35, no. 2, pp. 153–158, 1997, doi: 10.1016/S0005-7967(96)00085-X.

[29] M. Azmandian, M. Hancock, H. Benko, E. Ofek and A. D. Wilson, "Haptic Retargeting: Dynamic Repurposing of Passive Haptics for Enhanced Virtual Reality Experiences," *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 1968–1979, 2016, doi:10.1145/2858036.2858226.

[30] P. Willemsen, A. A. Gooch, W. B. Thompson and S. H. Creem-Regehr, "Effects of Stereo Viewing Conditions on Distance Perception in Virtual Environments," *Presence*, vol. 17, no. 1, pp. 91-101, 2008, doi:10.1162/pres.17.1.91.

[31] E. Ebrahimi, S. V. Babu, C. C. Pagano and S. Joerg, "Towards a comparative evaluation of visually guided physical reach motions during 3D interactions in real and virtual environments," *2016 IEEE Symposium on 3D User Interfaces (3DUI)*, Greenville, SC, 2016, pp. 237-238.

[32] E. Ebrahimi, "Investigating Embodied Interaction in Near-Field Perception-Action Re-Calibration on Performance in Immersive Virtual Environments," PhD dissertation, School of Computing, Clemson University, South Carolina, 2017.

[33] B. Lok, S. Naik, M. Whitton and F. P. Brooks, "Effects of handling real objects and self-avatar fidelity on cognitive task performance and sense of presence in virtual environments," *Presence: Teleoperators and Virtual Environments*, vol. 12, no. 6, pp. 615-628, 2003.

[34] J.-L. Rodríguez, R. Velázquez, C. Del-Valle-Soto, S. Gutiérrez, J. Varona and J. Enríquez-Zarate, "Active and Passive Haptic Perception of Shape: Passive Haptics Can Support Navigation," *Electronics*, vol. 8, no. 3, 355, 2019, doi: 10.3390/electronics8030355.

[35] R. Velázquez, E. Pissaloux, C. Del-Valle-Soto, M. Arai, L.J. Valdivia, J.A. Del Puerto-Flores and C.A. Gutiérrez, "Performance Evaluation of Active and Passive Haptic Feedback in Shape Perception," *IEEE 39th Central America and Panama Convention*, pp. 1-6, 2019, doi: 10.1109/CONCAPANXXXIX47272.2019.8977077.

[36] H.N. Schifferstein and M.P. Cleiren. "Capturing Product Experiences: A Split-Modality Approach," *Acta Psychologica*, vol. 118, no. 3, pp. 293-318, 2005, doi: 10.1016/j.actpsy.2004.10.009.

[37] C. Luo, Y. Shen and Y. Liu, "Look and Feel: The Importance of Sensory Feedback in Virtual Product Experience," *Proceedings of Fortieth International Conference on Information Systems*, pp. 1-9, 2019.

[38] R.S. Elder and A. Krishna. "The "Visual Depiction Effect" in Advertising: Facilitating Embodied Mental Simulation through Product Orientation," *Journal of Consumer Research*, vol. 38, no. 6, pp. 988-1003, 2012, doi:10.1086/661531.

[39] A. Krishna, "An Integrative Review of Sensory Marketing: Engaging the Senses to Affect Perception, Judgment and Behavior," *Journal of Consumer Psychology*, vol. 22, no. 3, pp. 332-351, 2012, doi:10.1016/j.jcps.2011.08.003.

[40] A.E. Schlosser, "Experiencing Products in the Virtual World: The Role of Goal and Imagery in Influencing Attitudes Versus Purchase Intentions," *Journal of Consumer Research*, vol. 30, no. 2, pp. 184-198, 2003, doi: 10.1086/376807.

[41] L.W. Barsalou, "Grounded Cognition," *Annual Review of Psychology*, vol. 59, no. 1), pp. 617-645, 2008, doi: 10.1146/annurev.psych.59.103006.093639.

[42] N. Schwarz, Feelings-as-Information Theory, *Handbook of Theories of Social Psychology*, P.A.M.V. Lange, A. Kruglanski and E.T. Higgins, eds., Thousand Oaks, CA: Sage, pp. 289-308, 2012.

[43] A. Krishna and N. Schwarz, "Sensory Marketing, Embodiment, and Grounded Cognition," *Journal of Consumer Psychology*, vol. 24, no. 2, pp. 159-168, 2014, doi:10.1016/j.jcps.2013.12.006.

[44] J. Peck and T.L. Childers, "To Have and to Hold: The Influence of Haptic Information on Product Judgments," *Journal of Marketing*, vol. 67, no. 2, pp. 35-48, 2003b, doi:10.1509/jmkg.67.2.35.18612.

[45] D. Löffler, L. Arlt, T. Toriizuka, R. Tscharn and J. Hurtienne, "Substituting Color for Haptic Attributes in Conceptual Metaphors for Tangible Interaction Design," *Proceedings of the TEI '16: Tenth International Conference on Tangible, Embedded, and Embodied Interaction (TEI '16)*, pp. 118–125, 2016, doi:10.1145/2839462.2839485.

[46] D. Löffler, R. Tscharn and J. Hurtienne, "Multimodal Effects of Color and Haptics on Intuitive Interaction with Tangible User Interfaces," *Proceedings of the Twelfth International Conference on Tangible, Embedded, and Embodied Interaction (TEI '18)*, pp. 647–655, 2018, doi:10.1145/3173225.3173257.

[47] L. H. Kim and S. Follmer, 2019. "SwarmHaptics: Haptic Display with Swarm Robots," *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*, pp. 1-13, 2019, doi:10.1145/3290605.3300918.

[48] C. Carvalheiro, R. Nóbrega, H. da Silva and R. Rodrigues, "User Redirection and Direct Haptics in Virtual Environments," *Proceedings of the 24th ACM international conference on Multimedia (MM '16)*, pp. 1146–1155, 2016, doi:10.1145/2964284.2964293.

[49] A. Lassagne, A. Kemeny, J. Posselt and F. Merienne, "Performance Evaluation of Passive Haptic Feedback for Tactile HMI Design in CAVEs," *IEEE Transactions on Haptics*, vol. 11, no. 1, pp.119-127, 2018, doi: 10.1109/toh.2017.2755653.

[50] R.D. Joyce and S.K. Robinson, "Passive Haptics to Enhance Virtual Reality Simulations," *AIAA Modeling and Simulation Technologies Conference*, 2017, doi:10.2514/6.2017-1313.

[51] G. Park, S. Choi, K. Hwang, S. Kim, J. Sa and M. Joung, "Tactile effect design and evaluation for virtual buttons on a mobile device touchscreen," *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI '11)*, pp. 11–20, 2011, doi:10.1145/2037373.2037376.

[52] S. Dabic, J. Navarro, J. M. Tissot and R. Versace, "User perceptions and evaluations of short vibrotactile feedback," *Journal of Cognitive Psychology*, vol. 25, no. 3, pp. 299-308, 2013, doi:10.1080/20445911.2013.768997.

[53] L. Diwischek and J. Lisseman, "Tactile feedback for virtual automotive steering wheel switches," *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '15)*, pp. 31–38, 2015, doi:10.1145/2799250.2799271.

[54] L.-P. Cheng, L. Chang, S. Marwecki and P. Baudisch, "ITurk: Turning Passive Haptics into Active Haptics by Making Users Reconfigure Props in Virtual Reality," *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*, pp. 1–10, 2018, doi:10.1145/3173574.3173663.

[55] J. C. McClelland, R. J. Teather and A. Girouard, "Haptobend: shape-changing passive haptic feedback in virtual reality," *Proceedings of the 5th Symposium on Spatial User Interaction (SUI '17)*, pp. 82–90, 2017, doi:10.1145/3131277.3132179.

[56] J. Ye, S. Badiyani, V. Raja and T. Schlege, "Applications of Virtual Reality in Product Design Evaluation," *Human-Computer Interaction. HCI Applications and Services*, LNCS 4553, J.A. Jacko, ed. Springer, 2007, pp. 1190–1199.

[57] M. G. Violante, F. Marcolin, E. Vezzetti, F. Nonis and S. Moos (2019), "Emotional Design and Virtual Reality in Product Lifecycle Management (PLM)," in Sustainable Design and Manufacturing 2019. KES-SDM 2019. Smart Innovation, Systems and Technologies, Singapore: Springer, pp. 177-187, 2019.

[58] C. S. Falcao and M. M. Soares, "Applications of Haptic Devices & Virtual Reality in Consumer Products Usability Evaluation," *Advances in Ergonomics In Design, Usability & Special Populations: Part I*, AHFE Conference, pp. 377-383, 2014.

[59] C.E. Osgood, G.J. Suci and P.H. Tannenbaum, *The measurement of meaning*, Champaign, IL: University of Illinois press, 1957.

[60] M. Perez, S. Ahmed-Kristensen, P. Brunn and H. Yanagisawa, "Investigating the influence of product perception and geometric features," *Research in Engineering Design*, vol. 28, no. 3, pp. 357-379, 2017, doi:10.1007/s00163-016-0244-1.

[61] S. Mondragón, P. Company and M. Vergara, "Semantic differential applied to the evaluation of machine tool design," *International Journal of Industrial Ergonomics*, vol. 35, no. 11, pp. 1021-1029, 2005, doi:10.1016/j.ergon.2005.05.001.

[62] S.W. Hsiao, F.Y. Chiu and C.S. Chen, "Applying aesthetics measurement to product design," *International Journal of Industrial Ergonomics*, vol. 38, no. 11-12, pp. 910-920, 2008, doi:10.1016/j.ergon.2008.02.009.

[63] J. Al-Hindawe, "Considerations when constructing a semantic differential scale," *La Trobe papers in linguistics*, vol. 9, no. 7, pp. 1-9, 1996.

[64] S. Achiche, A. Maier, K. Milanova and A. Vadean, "Visual Product Evaluation: Using the Semantic Differential to Investigate the Influence of Basic Geometry on User Perception," *ASME 2014 International Mechanical Engineering Congress and Exposition*, pp. V011T14A056-V011T14A056, Nov. 2014.

[65] F. Felip, J. Galán, C. García-García and E. Mulet, "Influence of presentation means on industrial product evaluations with potential users: a first study by comparing tangible virtual reality and presenting a product in a real setting," *Virtual Reality*, 2019, doi:10.1007/s10055-019-00406-9.

[66] F. Faul, E. Erdfelder, A.-G. Lang and Buchner, A., "G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences," *Behavior Research Methods*, vol. 39, pp. 175-191, 2007, doi:10.3758/BF03193146.

[67] T. Van Hecke, "Power study of ANOVA versus Kruskal-Wallis test," *Journal of Statistics and Management Systems*, vol. 15, no. 2-3, pp. 241-247, 2012, doi:10.1080/09720510.2012.10701623.

Julia Galán

Julia Galán holds a PhD in Fine Arts (Universitat Politècnica de València, Spain, 1991), has been a Professor at Universitat Jaume I since 1994 and is the director of the DACTIC research group since 2010. Her research interests focus on the interconnections between design and contemporary art that occur mainly through the use of new technologies to innovate, and in the field of product presentation, studying how the means of presentation influence the user's perception of the product.

Carlos García-García

Carlos García-García earned a BSc in Ceramics (Escuela Superior de Cerámica de L'Alcora, Spain, 2009), an MSc in Professor of Secondary Education (Univesitat Jaume I, Spain, 2011), an MSc in Design and Manufacturing (Univesitat Jaume I, Spain, 2012) and a PhD in Industrial Technologies (Univesitat Jaume I, Spain, 2014). He is an associate professor with the Department of Industrial Systems Engineering and Design of Universitat Jaume I. His main research focuses on the development and management of co-creative methodologies within the fields of Art and Design, and the field of product presentation, studying how the means of presentation influences the user's perception of the product.

Francisco Felip

Francisco Felip received the BS degree in Industrial Design from Universitat Jaume I (Spain, 1997) and the PhD degree in Fine Arts from Universitat Politècnica de València (Spain, 2008). He is an associate professor with the Department of Industrial Systems Engineering and Design of Universitat Jaume I. His research focuses on the border area between Art and Design, studying the creative synergies that occur between them. His current research focuses on the field of product presentation, studying how the means of presentation influences the user's perception of the product.

Manuel Contero

Manuel Contero is a Full Professor of Engineering Graphics and CAD with the Graphic Engineering Department at the Universitat Politècnica de València, Spain (UPV). He earned an MSc degree in Electrical Engineering in 1990 (Univesitat Jaume I, Spain) and a PhD in Industrial Engineering in 1995 (Univesitat Jaume I, Spain). In 1993 he joined Universitat Jaume I as assistant professor, promoting to associate professor in 1997. In 2000 he returned to UPV, being appointed full professor in 2008. His research interests focus on sketch-based modeling, collaborative engineering, human computer interaction, development of spatial abilities, and technology enhanced learning.

# Motivic Pattern Classification of Music Audio Signals Combining Residual and LSTM Networks

Aitor Arronte Alvarez[1,2]*, Francisco Gómez[1]

[1] Universidad Politécnica de Madrid, Madrid (Spain)
[2] University of Hawaii at Manoa, Honolulu (USA)

Received 13 August 2020 | Accepted 14 January 2021 | Published 21 January 2021

## Abstract

Motivic pattern classification from music audio recordings is a challenging task. More so in the case of a cappella flamenco *cantes*, characterized by complex melodic variations, pitch instability, timbre changes, extreme vibrato oscillations, microtonal ornamentations, and noisy conditions of the recordings. Convolutional Neural Networks (CNN) have proven to be very effective algorithms in image classification. Recent work in large-scale audio classification has shown that CNN architectures, originally developed for image problems, can be applied successfully to audio event recognition and classification with little or no modifications to the networks. In this paper, CNN architectures are tested in a more nuanced problem: flamenco *cantes* intra-style classification using small motivic patterns. A new architecture is proposed that uses the advantages of residual CNN as feature extractors, and a bidirectional LSTM layer to exploit the sequential nature of musical audio data. We present a full end-to-end pipeline for audio music classification that includes a sequential pattern mining technique and a contour simplification method to extract relevant motifs from audio recordings. Mel-spectrograms of the extracted motifs are then used as the input for the different architectures tested. We investigate the usefulness of motivic patterns for the automatic classification of music recordings and the effect of the length of the audio and corpus size on the overall classification accuracy. Results show a relative accuracy improvement of up to 20.4% when CNN architectures are trained using acoustic representations from motivic patterns.

## Keywords

Motivic Patterns, Convolutional Neural Networks, Data Augmentation, Audio Signal Processing, Music Information Retrieval.

DOI: 10.9781/ijimai.2021.01.003

## I. Introduction

THE automatic extraction, discovery, and classification of motivic patterns from music audio recordings is a task that has gathered the attention of the Artificial Intelligence community in general, and the Music Information Retrieval (MIR) community in particular [1], [2], [3]. Repeated melodic patterns are important in the analysis and understanding of music. More recently, research has shown that repeated small musical patterns that are transformed up to a certain extent, play an important role in establishing music similarity in orally transmitted songs [4].

The computational study of orally transmitted vocal music repertoires present different types of problems associated with the audio signal obtained from such recordings. The high degree of variability in the audio signal has to do with the environmental conditions of the recordings, the improvisatory nature of the singing styles, and the rapid fluctuation of wide vibrato ranges. In flamenco music, these difficulties are even more acute, since intervals are often smaller than the half-tone. A cappella flamenco *cantes* exhibit characteristic melodic features such as conjunct degrees in the melodic movement, high degree of ornamentation, extreme pitch oscillations, microtonal variation, and constant timbre changes. These features make the automatic extraction of motivic patterns from audio recordings an especially challenging task.

The computational study of flamenco music has concentrated on the melodic characterization of *cantes* [5], [6], melodic pattern extraction [2], and the modelling of melodic variation [7]. Pattern extraction methods in *flamenco* research have used humans to extract relevant segments and melodic motifs [2], [5]. To our knowledge, exclusively data-driven approaches for the automatic intra-style classification of music audio signals have not yet been developed in previous research.

Different approaches in the MIR research literature have considered the use of Convolutional Neural Networks (CNN) for music tagging, genre prediction, and music classification. CNN have been used for mood and genre prediction using mel-spectrograms as the input representation [8]; the classes used in this study include genres (classical and pop), and moods (soft, ambient) among other label descriptors. Image classification CNN architectures were used for music classification based on general music style tags [9]. Other transfer learning approaches on MIR tasks include multi-label classification and prediction [10], and general-purpose music classification [11]. In the audio signal processing research in general, CNN architectures were used on large-scale audio event classification [12], showing that image architectures can be reused for audio processing task with some adjustments in the architectures' filter size.

Other applications of deep neural networks to music analysis and its computational understanding include low-level tasks such

* Corresponding author.
E-mail address: arronte@hawaii.edu

as beat tracking [13], onset detection [14], tempo estimation [15], and chord recognition [16]. These low-level tasks attempt to learn representations of acoustic phenomena directly from the audio signal. Higher-level tasks learn representations that can map acoustic features into more abstract musical concepts such as music style classification [17], and singer identification [18] amongst others. MIR applications of high-level music tasks strongly depend on pre-existing knowledge and domain adaptation. In the approach presented in this article, no hand-crafting or domain adaptation is needed, since motivic patterns are extracted directly from the audio signal without prior knowledge.

This paper investigates the usefulness of motivic patterns for the automatic classification of different styles of flamenco music by using different CNN architectures originally conceived for image classification tasks. This research also extends the computational study of motivic patterns in flamenco music by presenting a pipeline for motivic contour extraction from audio recordings based on an approximation scheme. Then classification task is performed from the patterns obtained using the log mel-spectrograms extracted from the recordings' raw audio signals by using different CNN. The contributions of this research are the following: 1) We propose a motivic extraction pipeline as a preprocess step, which improves the classification accuracy of all the architectures tested. 2) It is shown that CNN architectures from very different domains can achieve competitive results with state-of-the-art algorithms while simplifying the learning process and making it computationally more efficient, mostly because of the pipeline introduced in this article. 3) A neural architecture is presented that is able to use some of the advantages of image classification CNN models, particularly as audio feature extractors, while at the same time adding recurrent layers with bidirectional LSTMs that are able to process musically relevant sequential data, adding more explanatory power to the results. 4) We make code and data of the experiments publicly available1.

The different sections of this article are organized as follows: Section II presents the corpus of flamenco recordings (COFLA) and describes its contents and music characteristics. Section III sets forth the motivic pattern extraction method and audio features used as the input of the different architectures. Section IV describes the CNN models used as baselines in this research and the hybrid recurrent model introduced in this article. Section V presents the experiments and data used to test the different CNN architectures. Section VI outlines the results of the experiments and discusses the main findings, improvements, and shortcomings. Section VII concludes by listing the main contributions of this research and possible future lines of work.

## II. Corpus of Flamenco Recordings

Flamenco is an orally transmitted musical tradition from Andalusia, a region in the south of Spain. Its rich history and musical characteristics are derived from the region's cultural exchanges amongst various populations over centuries, most notably Andalusian-Romani, Jews, and Arabs. Some of the key characteristics of flamenco music such as pitch instability, the use of intervals smaller than the half-tone, the amount of variation from phrase to phrase and from singer to singer, are derived from its improvisatory nature. Even though improvisation plays a very important role in the conception of flamenco music, it is a highly structured and elaborated musical tradition [19].

Flamenco music centers around the singing voice usually accompanied by guitar, hand-clapping, and other percussion instruments like the *cajón*. Melodies are characterized by a combination of short and long notes with syllabic ornamentations (melismas), that are placed in specific locations in a phrase [20]. Flamenco singers learn melodies belonging to different styles and acquire singing techniques by oral transmission.

The main focus in the computational study of flamenco music is the development of algorithms that target the analysis of the singing voice [19]. Flamenco music, like most orally transmitted musical cultures, lacks music transcriptions of the repertoire. For that reason, corpora of audio recordings, with their corresponding meta-data, are the main source of research data. In this article corpus COFLA is used [20]. The corpus consists of more than 1,800 music recordings taken from flamenco anthologies. This corpus follows the research corpora principles formulated by Serra [21]. The main characteristics of the corpus, as summarized by its authors [20], are:

- Exhaustiveness: the corpus is composed of all anthologies published on CD during the 20th century, and are considered references for music critics and musicologists.
- Representation: each anthology represents a wide variety of styles and their variants.
- Sound quality: the audio quality varies greatly amongst recordings, but all recordings comply with a minimum standard.
- Commercial availability: all recordings are available to the general public, which facilitates the acquisition and allows for the establishment of ground truth data.

In this research, the following styles and substyles are used from the corpus COFLA: *tonás* (*deblas, martinetes, and saetas*), and *fandangos*. Stylistically, the *tonás* is an important group of a capella *cantes* sung in free rhythm, where singers choose their own reference pitch and perform variations on a given melody. A *toná* normally is composed of four verses of eight syllables each. Tempo is not strictly kept during a single piece and ornamentation is heavily used by singers. In the *tonás* style, *deblas* are characterized by melismatic ornamentations with more abrupt changes than the rest of the compositions in the *tonás* style. *Martinetes,* also a *toná* variant, differ slightly in its melodic model from the *debla* and, even though it is mostly sung without accompaniment, it uses a hammer and anvil as percussion instruments. *Saetas,* another *toná* variant, have a religious content in its lyrics and is stylistically closer to the *debla* in its usage of long and sustained notes combined with melismatic ornamentations. The style of *fandango* is more differentiated from the variants in the *tonás*. A *fandango* is a musical style associated with a dance and is rhythmically more complex than the *tonás*.

We select a sample from the corpus COFLA consisting of 13 *deblas,* 12 *saetas,* and 50 *martinetes*. The *martinete* subsample contains a wider variety of singing styles, and to some researchers it can be decomposed into 2 subtypes [22]. The current sample presents different stylistic challenges and difficulties for the automatic classification of motifs based on their substyle. First of all, 3 of the classes belong to the same genre (*deblas, martinetes, and saetas*), which means that these substyles share more musical traits with each other than with the *fandango*. This will add another level of complexity to the computational analysis, considering that previous studies have dealt only with the classification of different genres of music. In this paper the analysis is restricted to a very specific genre of music, namely flamenco, but also it is restricted to unaccompanied vocal music of different subgenres of flamenco.

## III. Audio Features and Contour Extraction Method

From the sample of songs described in Section II, we extract musical motifs following a pipeline based on two main components: a contour simplification method and the BIDE pattern mining algorithm [18]. The purpose of this pipeline is to extract statistically and musically relevant motifs from flamenco audio recordings characterized by high

[1] https://rb.gy/q3ppg0

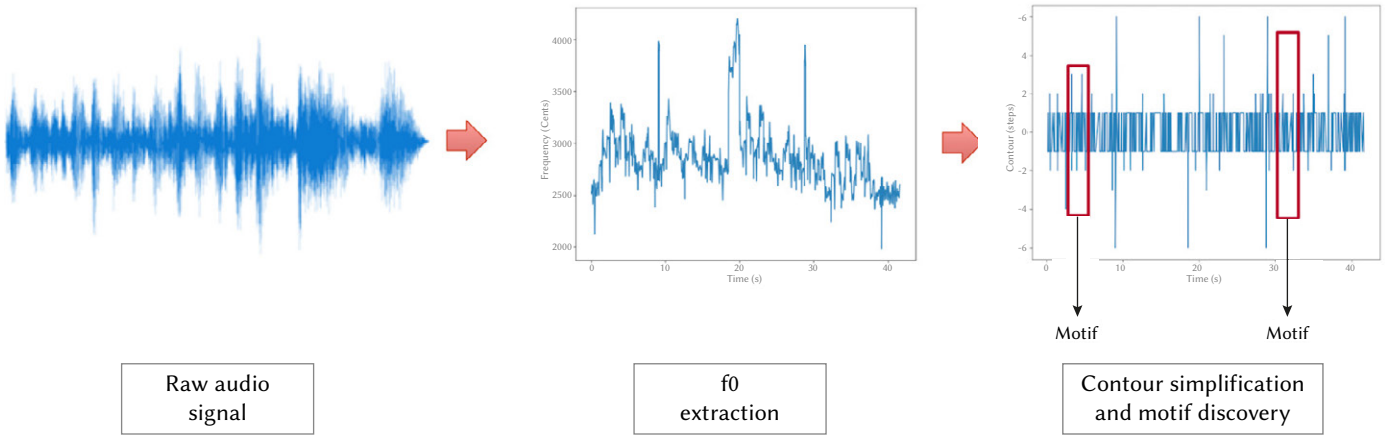



Fig. 1. Motivic pipeline for the extraction of patterns from raw audio signals.

instability of pitch. The pipeline here described attempts to solve the problem of reducing the pitch variability in the audio signal, with an approximation method that uses ranges of pitch distances instead of a fully tempered system such as the one used in western classical music. The steps of this motivic pipeline, as shown in Fig. 1, are the following:

1. Extract the fundamental frequency $f_0$ from the audio signals of the songs
2. Apply a melodic contour simplification function $C(f_0)$ based on the extracted $f_0$ for each one of the songs, thus obtaining an approximation of $f_0$
3. Apply the BIDE algorithm on the melodic contours obtaining a dictionary of motifs for the entire collection
4. Generate log mel-spectrograms for each motif in the dictionary

The fundamental frequency is extracted from raw audio signals by using a sinusoid extraction and salience function [24]. A sampling rate of 44.1 KHz and a window step of 256 samples are used. Then, a melodic contour simplification procedure is performed to extract meaningful motivic representations of a cappella flamenco *cantes* from the fundamental frequency. In previous studies [6], contour simplification procedures have been used to obtain consistent representations of flamenco melodic segments by converting complex pitch fluctuations to equal-step segments. Since we are studying motivic patterns in complex flamenco vocal pieces, we are interested in exploring the unequal microtonal nature of this type of music. In order to accomplish this goal, our contour simplification process takes into account ranges of cent-based distances instead of set of pitches as presented in previous work [6].

We follow these steps to find a curve approximation to $f_0$ given a step length of ε=66 cents based on previous approximation approaches [20]:

- Given a set of points $P$ in $f_0$ we say that a line segment $L$ is bounded by all points in $P$ given a maximum accepted step size of ε.
- The output of this procedure is a contour simplification function of *f0*.

Once this output is computed, a contour $C$ is obtained based on the following distance specification in cents:

- If the distance $d$ between two points <=66 then, $d$=1
- If the distance $d$ between two points >66 or $d$ <=132 then, $d$=2
- If the distance $d$ between two points >132 <=198 then, $d$=3
- 4 otherwise

The result is a vector of contour points represented in the time domain. The signs + and – are used to specify whether the direction of the contour ascends (+) or descends (-). Sudden jumps in frequency are eliminated due to external noise conditions.

Once the approximation function is created, the BIDE algorithm is used to discover motifs in the contour sequences. Motifs that are repeated at least 3 times in a single song are kept. From the dictionary of motifs, log mel-spectrograms are computed from the 2D time-frequency motivic patches, with hop and window sizes of 25 ms. The input size for all samples is 128x426, zero-padding smaller audio files.

## IV. Baseline and Hybrid Architectures

Transfer learning approaches in deep neural networks have shown to be not only computationally more efficient in achieving competitive results, but also show how representations from one task can be transferred to another task. The different CNN architectures developed initially for image classification problems and used in this article's experimental study are, *DenseNet-161,* and *ResNet-50.* A state-of-the-art Convolutional Recurrent Neural Network (CRNN) architecture developed specifically for music classification is also used as a baseline [26]. Filter sizes and strides are kept small, 3x3 and 1x1 respectively. This is mostly because of the small size of the audio input.

### A. ResNet-50

Deep residual networks were conceived to address the problem of learning degradation in deep nets. Residual networks are based on the idea of stacking layers and an underlying mapping that is optimized [27]. The model used in this study, *Resnet-50,* is transformed in a similar way as in [12] by removing the stride of 2x2 in the first convolutional layer, and reducing the size of the first convolutional filter from 7x7 to 3x3. In addition to that, and in order to maintain the input tensor size of the mel-spectrogram and to leave the *ResNet-50* architecture intact for baseline purposes, we add an initial convolutional layer with filters of size 3x3 and stride of 1.

### B. DenseNet-161

CNN that have shorter connections between layers that are closer to the input and output of a network have shown to be more accurate. This paradigm is followed by the *DenseNet* model [28]. We make the same modifications to the architecture as in *ResNet-50.*

### C. CRNN

A model for music audio tagging that has shown state-of-the-art results is the CRNN of Choi et al. [8]. This model utilizes the benefits of CNN as feature extractors and the sequential characteristics of Recurrent Neural Networks (RNN) to summarize time-dependent data as the one obtained from musical pieces.

### D. Hybrid Recurrent Architecture

In this work we attempt to use and exploit the advantages of CNN layers as feature extractors and add recurrent components in the last layers to capture sequential characteristics present in music audio data.

Deep learning architectures for audio classification are normally divided into front-end and back-end components [30]. The front-end, is the part of the model that tries to learn a representation based on the input signal. The back-end is in charge of predicting a given output based on the representation obtained in the front-end. Our hybrid model uses shallow residual blocks present in *Resnets* as a front-end, and a recurrent neural model as a back-end. The overall goal of this architecture is to simplify the already high cost of deep learning methods, especially in the front-end, while trying to improve state-of-the-art results by adding domain specific knowledge in the back-end. We try therefore to reduce the number of parameters of very deep networks by adding recurrent layers.

The shallow *residual* network proposed is composed of only 2 residual blocks, which reduces the computational cost and overall training time when compared to denser *Resnet* models normally utilized in the computer vision literature. Small filters of 3x3 with a stride of 1 are used in all convolutional layers to capture local feature-maps, and finer low-level spectral features. As the back-end two-stacked Bidirectional Long Short-Term Memory (BLSTM) layers to capture longer, time-dependent, features [31]. Fig. 2 presents a high-level overview of the architecture described.

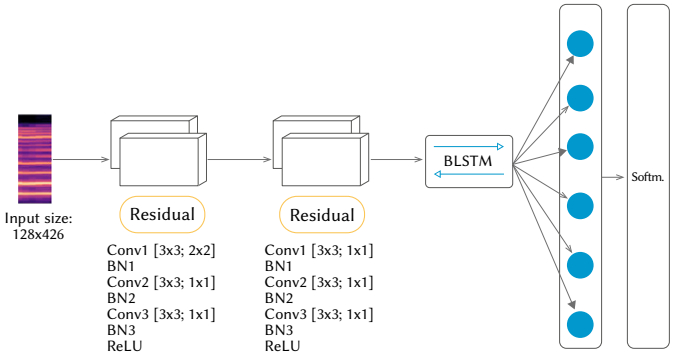



Fig. 2. Neural network architecture overview. Residual blocks contain convolutional layer dimensions (filter size, and stride), and batch size normalization (BN) and ReLU components.

In our back-end, the BLSTM is used to process in both directions (forward and backwards) the embedding obtained from the residual layers. The output of this layer will be a high level, vector representation of the time-dependent features of the motifs. The sequential operation done by the BLTSM can be represented as an input sequence $x = \{x_1, \dots, x_T\}$ that produces an output sequence $y = \{y_1, \dots, y_T\}$ where the input $x$ is a vector of acoustic features at the frame level. A BLTSM is composed of a forward and backward LSTM, where the forward LSTM $\vec{f}$ reads the input sequence as it is ordered, and estimates the forward hidden states $\vec{h}_1, \dots, \vec{h}_T$ from $t = 1$ to $T$. The backward LSTM $\overleftarrow{f}$ computes the sequence in reverse order obtaining the backward hidden states iterating back from $t = T$ to 1:

$$\vec{h}_t = H(W_{x\vec{h}} x_t + W_{\vec{h}\vec{h}} \vec{h}_{t-1} + b_{\vec{h}}) \quad (1)$$

$$\overleftarrow{h}_t = H(W_{x\overleftarrow{h}} x_t + W_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t-1} + b_{\overleftarrow{h}}) \quad (2)$$

$$y_t = W_{\vec{h}y} \vec{h}_t + W_{\overleftarrow{h}y} \overleftarrow{h}_t + b_y \quad (3)$$

where $H$ is the hidden layer function, $W$ the matrix of weights, and $b$ the bias vector.

The final layer of the architecture presented is a fully connected neural network (FCNN) layer with a softmax function to classify the sequences according to the style label. Fig. 3 shows a high-level motivic pipeline overview.

## V. Experimental Methodology

We compare all models in the sub-style classification of musical patterns extracted from corpus COFLA, as described in Section III, and use 2D log mel-spectrograms as the input of the networks. The dataset used in this study is composed of 111,076 audio motifs extracted from the 4 sub-collections. We noted in the initial stages of the study that extremely short motifs (<0.5 seconds) do not help in the classification accuracy; for that reason only motifs that are >= 0.5 seconds in duration are kept. This resulted in a corpus of only 10,640 motifs, of which 1,573 were obtained from the *debla* sub-style, 129 from the *fandango*, 5,027 from the *martinete*, and 3, 915 from the *saeta*. We can see how certain sub-styles are richer in motivic patterns than others, and note that the *fandango* sub-collection in particular, is much less varied in longer motivic patterns (>= 0.5 seconds). This unbalanced dataset allows us to test data augmentation techniques in the context of audio musical data.

Unlike previous approaches to music classification and tagging in MIR, the approach presented will only learn a small segment of the entire audio signal. This segmentation based on the extraction of relevant motivic data will greatly benefit the representation learned, and reduce the total training time. From an information-theoretic stand point, it can be argued that reducing the amount of irrelevant information to the task will act as an implicit optimizer for the neural architectures, while at the same time obtaining more explainable results in music terms.

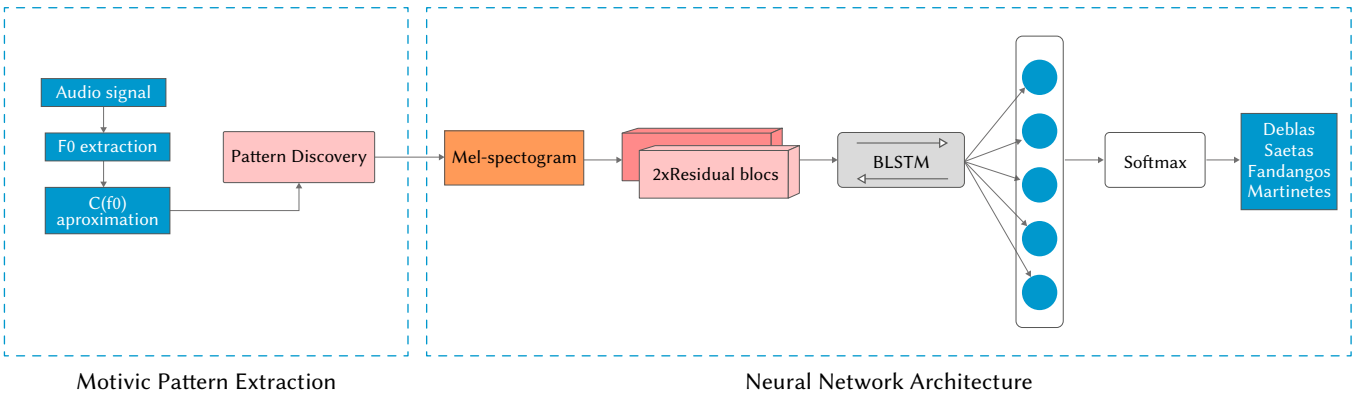



Fig. 3. High-level motivic pipeline overview.

### A. Data Augmentation

In this experiment a recent method for data augmentation developed for Automatic Speech Recognition (ASR) called SpecAugment is used [32]. Instead of producing deformations to the raw audio signal like other audio-based data augmentation techniques [33-34], SpecAugment operates directly on the spectrogram by warping it in the time direction, masking frequency channels, and masking blocks of utterances. The method follows a similar rationale as image data augmentation techniques.

We concentrate on the two augmentation policies that seem to be the most effective in ASR tasks [32]: frequency masking, and time masking. Frequency masking works on $m$ consecutive mel frequency channels $[m_0, m_0+m]$, where $m$ is chosen from a uniform distribution from 0 to the frequency mask parameter $M$. Time masking works in a similar way by applying the masking to $t$ consecutive time steps. We compare the two data augmentation policies with the original unbalanced dataset, and apply the following number of transformations by class:

- For *fandango* style a total of 8 augmentations per spectrogram is performed; 4 of time masking and 4 of frequency masking, resulting in a total subset of 1,032.
- For the rest of the styles we apply one of each augmentations in only 50% of their respective subsets. Resulting in 3,146 *deblas*, 10,054 *martinetes*, and 7,830 *saetas*.

The total dataset after augmentation contains 22,062 spectrograms of motifs. The comparative differences in motivic samples by sub-style are presented in Fig. 4.

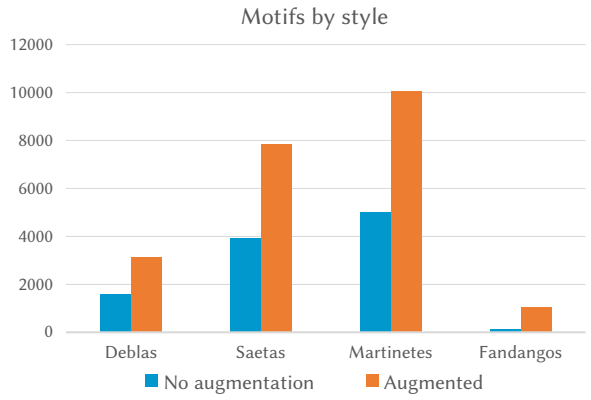



Fig. 4. Motivic patterns by sub-style.

### B. Training

All the models use the Adam optimizer [35] and data augmentation dynamically during training. We divide the data in training, and validation subsets, making the 60% and 20% respectively of the entire dataset, leaving the remaining 20% for testing. AUC-ROC, and accuracy scores are used to perform searches over the parameter space. It was found out during training that batch sizes of 20, and a total number of epochs of ~40 performed best in terms of accuracy and computational time. We found no sign of overfitting based on those measures in the validation subset, even in non-augmented sets.

Random initialization versus pretrained weights were also tested during training for all architectures. Results showed that pretrained weights not only perform better overall (>~2% accuracy) than random ones, but also decreased the training time (~10 epochs less to converge). We used pretrained weights from image classification tasks in our experiments.

## VI. Results and Discussion

Results in *Table I* show how the motivic pattern dataset has overall better results, with an average accuracy improvement of 13.1% across all models, with a maximum of 14.1% for *Resnet-50*, which indicates a relative improvement of 20.4%. Precision and recall measures also highlight the strength of motivic patterns for all models when compared to non-motivic data, and achieve an 85% precision for the proposed architecture with motivic patterns and no augmentation. These results shed light in the importance of motivic patterns in deep learning for music classification problems. This result can have significant implications in deep learning for MIR tasks, since shorter, more targeted audio data can significantly reduce the already huge computational costs of deep architectures. On the other hand, for multimedia systems in general, and MIR systems in particular, the effective retrieval of relevant audio information from big data can be improved with traditional sequential pattern mining techniques as a pre-step in the computational pipeline.

From a theoretical MIR point of view, our results highlight the importance of musically relevant features in deep learning systems as opposed to merely general audio features. In musically complex systems with melodic variability, microtonal ornamentations and contours, the extraction of relevant patterns can become a challenging task. The proposed contour simplification method takes into account small pitch fluctuations, and extracts small patterns (~0.5 seconds) that highlight particularities of a sub-style within flamenco music. These patterns may reveal vibrato styles, or ornamentation tendencies in singers for a particular style that may be difficult for the human ear to grasp. Further study should concentrate on the exploration of speech features combined with purely musical ones, which may aid the classification and automatic identification not only of styles, but singers as well.

The transferred architectures used in this study show how pretrained image weights can optimize the overall training procedure in music classification tasks and achieve competitive results with less training time. Since we are using mel-spectrograms of an audio signal as the input, the image-like 2-dimensional size of the input seems to be the reason why pretrained weights facilitate the accuracy results in less time when compared with random initialization. The hybrid architecture proposed outperforms the rest in terms of accuracy, AUC, precision, and recall. The performance values for non-motivic datasets with recurrent layers in the architecture indicates that these architectures can indirectly infer the temporal components of the data. Still the motivic dataset outperforms non-motivic ones for all models.

TABLE I. Model Results for the Motivic Patterns and Non-motivic Subsets

| *Model* | *Dataset* | *Accuracy* | *AUC* | *Prec.* | *Rec.* | *F1* |
|---|---|---|---|---|---|---|
| *Resnet-50* | Motivic | 0.832 | 0.894 | 0.801 | 0.769 | 0.785 |
| *Resnet-50* | Non-motivic | 0.691 | 0.792 | 0.631 | 0.617 | 0.624 |
| *Densenet-161* | Motivic | 0.817 | 0.881 | 0.735 | 0.731 | 0.733 |
| *Densenet-161* | Non-motivic | 0.683 | 0.769 | 0.61 | 0.589 | 0.599 |
| *CRNN* | Motivic | 0.821 | 0.886 | 0.78 | 0.757 | 0.768 |
| *CRNN* | Non-motivic | 0.796 | 0.853 | 0.714 | 0.711 | 0.712 |
| *ResLSTM* | Motivic | **0.911** | **0.91** | **0.848** | **0.813** | **0.83** |
| *ResLSTM* | Non-motivic | 0.824 | 0.882 | 0.816 | 0.79 | 0.803 |

The results in Table II show the data augmentation classification scores for the motivic pattern dataset. An accuracy improvement of 2.4% on the best model when using augmentation, highlights the importance of the data size in deep learning tasks. Since we obtain more than double of the original size from the motivic dataset, the improvements on the classification results seem to be logical.

SpecAugment, however, does not show an improvement as important as the one shown in the original study with speech data [32]. Further research should explore different ranges of masking parameters to determine the quality of the results and its appropriate use with musical vocal data.

TABLE II. Results for the Data Augmentation Policies Applied to the Motivic Pattern Dataset

| *Model* | *Augment.* | *Accuracy* | *AUC* | *Prec.* | *Rec.* | *F1* |
|---|---|---|---|---|---|---|
| *Resnet-50* | Frequency | 0.868 | 0.898 | 0.811 | 0.791 | 0.8 |
| *Resnet-50* | Time | 0.846 | 0.878 | 0.767 | 0.763 | 0.76 |
| *Densenet-161* | Frequency | 0.852 | 0.861 | 0.81 | 0.804 | 0.81 |
| *Densenet-161* | Time | 0.831 | 0.858 | 0.783 | 0.766 | 0.78 |
| *CRNN* | Frequency | 0.87 | 0.887 | 0.83 | 0.796 | 0.813 |
| *CRNN* | Time | 0.842 | 0.879 | 0.782 | 0.772 | 0.78 |
| *ResLSTM* | Frequency | 0.935 | 0.922 | 0.85 | 0.839 | 0.844 |
| *ResLSTM* | Time | 0.921 | 0.91 | 0.828 | 0.821 | 0.824 |

## VII. Conclusion

Overall the results indicate that the effect of motivic patterns in the classification accuracy of state-of-the-art CNN models is greater than the effect of data augmentation when using SpecAugment. Motivic patterns seem to provide important information in the classification of audio samples by style. Since CNN capture local-level features of a given audio sample, the utilization of motivic patterns seems to highlight higher level melodic features. Recurrent models on the other hand are less sensitive to non-motivic data. We also evaluated the importance of transfer learning in the context of musical audio data. The results of the transferred models are consistent with a recent large-scale audio classification study [12], which also extends the findings to music audio data. We specifically noted the ability of the networks to converge up to a state-of-the-art competitive accuracy with less training when using pretrained weights from image classification tasks. The proposed neural architecture outperforms state-of-the-art CRNN for music classification by taking advantage of the long-term sequence processing that the BLSTM net does. By combining BLSTM with shallow residual blocks, we take advantage of the smaller number of parameters required, and less processing time, when compared with deeper *resnets*.

This study presents an important case of deep learning optimization for audio signal processing, by extracting smaller, more targeted audio samples, discarding irrelevant information from the signal and learning more robust representations. This approach can be particularly interesting for low-resource MIR applications. It can also be easily adapted to sound event recognition and identification, and to speech recognition tasks that have a strong acoustic component such as accent, emotion, and dialect identification.

## References


[1] Dannenberg, R. B., and Hu, N. "Pattern discovery techniques for music audio," *Journal of New Music Research*, vol. 32, no.2, pp. 153-163, 2003.

[2] Pikrakis, A., Gómez, F., Oramas, S., Díaz-Báñez, J. M., Mora, J., Escobar-Borrego, F., Gómez, E., and Salamon, J. "Tracking Melodic Patterns in Flamenco Singing by Analyzing Polyphonic Music Recordings," in *International Society for Music Information Retrieval Conference, ISMIR*, Porto, Portugal, 2012, pp. 421-426.

[3] Gulati, S., Serra, J., Ishwar, V., and Serra, X. "Mining melodic patterns in large audio collections of Indian art music," in *2014 Tenth International Conference on Signal-Image Technology and Internet-Based Systems*, Marrakech, Morocco, 2014, pp. 264-271.

[4] Volk, A., Haas, W. B., and Kranenburg, P. "Towards modelling variation in music as foundation for similarity," in *Proceedings of the 12th International Conference on Music Perception and Cognition and the 8th Triennial Conference of the European Society for the Cognitive Sciences of Music*, Thessaloniki, Greece, 2012, pp. 1085-1094.

[5] Mora, J., Gomez Martin, F., Gómez, E., Escobar-Borrego, F. J., and Díaz-Báñez, J. M. "Characterization and melodic similarity of a cappella flamenco cantes," in *International Society for Music Information Retrieval Conference, ISMIR*, Utrecht, The Netherlands, 2016, pp. 9-13.

[6] Kroher, N., and Díaz-Báñez, J. M. "Audio-based melody categorization: Exploring signal representations and evaluation strategies" *Computer Music Journal*, vol. 41, no. 4, pp. 64-82, 2018.

[7] Kroher, N., and Díaz-Báñez, J. M. "Modelling melodic variation and extracting melodic templates from flamenco singing performances," *Journal of Mathematics and Music*, vol. 13, no. 2, pp. 150-170, 2019.

[8] Choi, K., Fazekas, G., and Sandler, M. "Automatic tagging using deep convolutional neural networks," arXiv preprint arXiv:1606.00298.

[9] Kim, T., Lee, J., and Nam, J. "Sample-level CNN architectures for music auto-tagging using raw waveforms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada, 2018, pp. 366-370.

[10] Dieleman, S., and Schrauwen, B. "End-to-end learning for music audio," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 6964-6968.

[11] Choi, K., Fazekas, G., and Sandler, M. "Transfer learning for music classification and regression tasks." arXiv preprint arXiv:1703.09179 2017.

[12] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., and Slaney, M. "CNN architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, 2017, pp. 131-135.

[13] Durand, S., Bello J. P., Bertrand D., and Gaël R. "Downbeat tracking with multiple features and deep neural networks," *in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015, pp. 409-413.

[14] Schlüter, J., and Böck, S. "Improved musical onset detection with convolutional neural networks," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, Florence, Italy, 2014, pp. 6979-6983.

[15] Corbera, F., and Serra, X. "Tempo estimation for music loops and a simple confidence measure," in *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR*, New York, USA, 2016, pp. 269-75.

[16] Korzeniowski, F., and Widmer, G. "A fully convolutional deep auditory model for musical chord recognition," in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, Salerno, Italy, 2016, pp. 1-6.

[17] Juhan, N., Choi, K., Lee, J., Chou, S., and Yang, Y. "Deep learning for audio-based music classification and tagging: Teaching computers to distinguish rock from bach," *IEEE signal processing magazine*, vol. 36, no. 1, pp. 41-51, 2018.

[18] Murthy, Y., Jeshventh, T. K. R., Zoeb, M., Saumyadip, M., and Shashidhar, G. K. "Singer identification from smaller snippets of audio clips using acoustic features and DNNs," in *2018 Eleventh International Conference on Contemporary Computing (IC3)*, Noida, India, 2018, pp. 1-6.

[19] Gómez, F., Dıaz-Bánez, J. M., Gómez, E., and Mora, J. "Flamenco music and its computational study," in *Mathematical Music Theory: Algebraic, Geometric, Combinatorial, Topological and Applied Approaches to Understanding Musical Phenomena*, World Scientific Publishing, Singapore, ch. 8, pp. 303-315.

[20] Kroher, N., Díaz-Báñez, J. M., Mora, J., and Gómez, E. "Corpus COFLA: a research corpus for the computational study of flamenco music," *Journal on Computing and Cultural Heritage (JOCCH)*, vol. 9, no. 2, pp. 1-21. 2016.

[21] Serra, X. "Creating research corpora for the computational study of music: the case of the Compmusic project," in *Audio engineering society conference: 53rd international conference: Semantic audio*, London, UK, 2014, article number 1-1, [9p.].

[22] Mora, J., Gómez, F., Gómez, E., and Díaz-Báñez, J. M. "Melodic contour and mid-level global features applied to the analysis of flamenco cantes," *Journal of New Music Research*, vol. 45, no. 2, pp. 145-159, 2016.

[23] H. Wang, J., and Han, J. "BIDE: Efficient mining of frequent closed sequences", in *Proceedings of the 20th international conference on data engineering*, Boston, MA, USA, 2004, pp. 79-90.

[24] J. Salamon, E. Gomez, and J. Bonada, "Sinusoid extraction and salience function design for predominant melody estimation," in *International Conference on Digital Audio Effects*, Paris, France, 2011, pp. 73–80.

[25] Díaz-Báñez, J. M., and A. Mesa. "Fitting Rectilinear Polygonal Curves to a Set of Points in the Plane.", in *European Journal of Operational Research* vol. 130, no. 1, pp. 214-222, 2001.

[26] Choi, K., Fazekas, G., Sandler, M., and Cho, K. "Convolutional recurrent neural networks for music classification", in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, US, 2017, pp. 2392-2396.

[27] He, K., Zhang, X., Ren, S., and Sun, J. "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, USA, 2016, pp. 770-778.

[28] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Honolulu, HI, USA, 2017, pp. 4700-4708.

[29] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, USA, 2016, pp. 2818-2826.

[30] Pons Puig, J., Nieto, O., Prockup, M., Schmidt, E. M., Ehmann, A. F., and Serra, X. "End-to-end learning for music audio tagging at scale," in *Proceedings of the 19th International Society for Music Information Retrieval Conference*, Paris, France, 2018, pp. 637-44.

[31] Graves, A., and Schmidhuber, J. "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol.18, no. 5-6, pp. 602-610, 2005.

[32] Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., and Le, Q. V. "Specaugment: A simple data augmentation method for automatic speech recognition". arXiv preprint arXiv:1904.08779. 2019.

[33] Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. "Audio augmentation for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, Dresden, Germany, 2015, pp. 3586-3589.

[34] McFee, B., Humphrey, E. J., and Bello, J. P. "A software framework for musical data augmentation," in *16th International Society for Music Information Retrieval Conference*, Malaga, Spain, 2015, pp. 248-254.

[35] Kingma, D. P., and Ba, J. "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

Aitor Arronte Alvarez

Aitor Arronte Alvarez is a machine learning researcher specializing in Music Information Retrieval, Audio Signal Processing, and Speech Recognition. He works at the University of Hawaii at Manoa at the Center for Language and Technology as a Technology Specialist. Aitor Arronte Alvarez holds a M. Eng. in Decision Systems Engineering and is currently finishing his Ph. D. at the Universidad Politécnica de Madrid.

Francisco Gómez

Francisco Gómez became Full Professor at Technical University of Madrid in 1994. He started doing research on computational geometry, computer graphics and facility location. In 2003 he switched to Music Information Retrieval and Computational Music Theory and has been doing research in this field since then. Francisco Gómez received a Ph.D. in Computer Science from the Technical University of Madrid under the supervision of Godfried Toussaint. His main interests in Music Information Retrieval and Computational Music Theory are music similarity, mathematical measures of rhythm complexity and syncopation, automated analysis of music traditions, especially flamenco music, Afro-Cuban music, Brazilian music and in general African music, teaching mathematics via the arts, and active learning methods in teaching mathematics. He has participated in several research projects funded by several Spanish agencies. Francisco Gómez teaches courses on Computer Graphics, Computational Geometry, Statistics, Algebra, Discrete Mathematics, and Pattern Recognition, among others. He uses innovative teaching methods such as inquiry-based teaching. In particular, he has used a collaborative version of the Moore method, Mazur's method for large audiences, and methods based on writing to teach mathematics.

# A Smart Collaborative Educational Game with Learning Analytics to Support English Vocabulary Teaching

Ahmed Tlili[1]*, Sarra Hattab[2], Fathi Essalmi[3], Nian-Shing Chen[4], Ronghuai Huang[1], Kinshuk[5], Maiga Chang[6], Daniel Burgos[7]*

[1] Smart Learning Institute of Beijing Normal University, Beijing (China)
[2] Higher Institute of Computer Science and Management of Kairouan, Kairouan (Tunisia)
[3] Management Information Systems Department, College of Business, University of Jeddah, Jeddah (Saudi Arabia)
[4] Department of Applied Foreign Languages, National Yunlin University of Science and Technology, 123 University Road, Section 3, Douliou, Yunlin 64002 (Taiwan)
[5] University of North Texas, 3940 N. Elm Street, G 150, Denton, TX, 76207 (USA)
[6] School of Computing and Information Systems, Athabasca University (Canada)
[7] UNIR iTED, Universidad Internacional de La Rioja (UNIR), Logroño (Spain)

Received 18 June 2020 | Accepted 17 February 2021 | Published 12 March 2021

## Abstract

Learning Analytics (LA) approaches have proved to be able to enhance learning process and learning performance. However, little is known about applying these approaches for second language acquisition using educational games. Therefore, this study applied LA approaches to design a smart collaborative educational game, to enhance primary school children learning English vocabularies. Specifically, the game provided dashboards to the teachers about their students in a real-time manner. A pilot experiment was conducted in a public primary school where the students' data from experimental and control groups, namely learning and motivation test scores, interview and observation, were collected and analyzed. The obtained results showed that the experimental group (who used the smart game with LA) had significantly higher motivation and performance for learning English vocabularies than the control group (who used the smart game without LA). The findings of this study can help researchers and practitioners incorporate LA in their educational games to help students enhance language acquisition.

## Keywords

Collaborative Learning, Data Analysis, Educational Games, Language Learning, Learning Analytics.

DOI: 10.9781/ijimai.2021.03.002

## I. Introduction

Arnold, Greenville and Doe [1] stated that the traditional methods for learning second languages are difficult and less engaged, resulting in negative learning outcomes. Flores [2] stated that Second Language Acquisition (SLA) strategies should be based on technologies since students of this era are technological natives. Additionally, immersive learning experiences play an important role in facilitating SLA. This immersion experience can be achieved by different technologies including games [3]. Several studies have proved that integrating playing and challenge while learning can improve students' outcomes [4]. Consequently, educational games have started gaining an increased attention from researchers and practitioners as a way of engaging students in learning. Educational games are games with the fundamental needs of learning by providing fun, motivation, creativity and social interaction [5], [6]. Especially, while playing games, students are situated in a gaming scenario to complete a series of learning tasks individually, collaboratively or even competitively. Despite that educational games are effective tools in enhancing the learning process, several research studies also reported that they are black boxes where teachers cannot unlock what students did in the learning process (except the final scores and levels cleared) and how they behaved towards the learning goal [7], [8].

Therefore, this study describes a smart collaborative educational game developed in this research, based on LA approaches, to teach primary school students English vocabulary. The game collected the students' learning interaction data and analyzed them to create dashboards that can help teachers understand how their students were learning using the game. The teachers can then provide the needed interventions to each student and each team accordingly. Additionally, this game adopted collaborative learning strategy in which achieving the game's goal depends on the efforts of all the team members. While LA approaches have been applied for several educational purposes, little attention has been paid to use these approaches for language acquisition [9], calling for further research in this regard. Additionally, despite that several educational games incorporated LA in the literature, to the best of our knowledge, none of these games applied collaborative learning while playing for second language acquisition.

* Corresponding author.
E-mail addresses: ahmed.tlili23@yahoo.com (A. Tlili), daniel.burgos@unir.net (D. Burgos).

Finally, this study compares the impact of the smart collaborative educational game and the non-smart version of the game on students' learning performance and motivation. The only difference between the smart and non-smart versions of the game is that the smart version incorporated LA to provide automatic dashboards about students' learning progress while the non-smart version did not.

The rest of the paper is structured as follows: Section II presents related work related to SLA and learning analytics in educational games. Section III describes the developed *Jungle animals* game- a smart collaborative educational game for teaching English vocabulary. Section IV presents the research method, while Section V presents the obtained results. Finally, Section VI discusses these results with the limitation of this study and future directions.

## II. Related Work

### A. Second Language Acquisition and Collaborative Learning

Second language acquisition is the learning and acquisition of a second language once the mother tongue or first language acquisition is established [10]. Hart and Risley [11] state that First Language Acquisition (FLA) is different than SLA because FLA occurs naturally and perhaps without any formal instruction, simply by students being constantly exposed to language rich environments over the course of many years. SLA, on the other hand, relies on more specific pedagogical approaches. In these settings, a major goal frequently is to formally teach students the elements of language that are learned much more informally in their native language.

Specifically, this research study aims to collaboratively teach English as a SLA. Collaborative learning is a pedagogical approach that implies students to work in groups to complete an activity or solve a given problem. Siemon, Becker, Eckardt, & Robra-Bissantz [12] introduce different principles for collaboration systems, as follows: "Reciprocity" refers to exchanging information and efforts. If a collaborator offers more effort, the other members have to return the effort when needed. The "common goal" is the most important factor in collaboration that motivates every team member to work with others. "Mutual respect and trust" enhance teamwork in a variety of ways. It is positively linked to a team performance. "Cohesiveness" refers to the perception of a team as one unified force. "Benevolence and commitment" mean that team members should not intentionally work against their team members and should deliver sufficient efforts to help their teams.

To teach SLA (individually or collaboratively), several researchers have used educational games to motivate students and provide learning environments and scenarios similar to real situations, as discussed in the next section.

### B. Educational Games and the Application of Learning Analytics

Several researchers have highlighted that educational games facilitate language learning since they are interactive and motivating. Additionally, they provide an environment similar to a real one, which enables students to easily practice the needed language and learn it effectively [3]. Surkamp and Viebrock [13] mentioned that games from simple vocabulary to role-playing games could enhance the language learning experience. However, Hung, Chang and Yeh [14] showed in a comprehensive literature review in SSCI language learning journals that only 4% of the published articles are related to Digital Game Based Language Learning (DGBLL). Additionally, the same authors reported, in their literature review, that 79% of DGBLL educators prefer to use off-the-shelf digital games over self-developed ones in order to reduce development cost and effort [14]. This shows the need to pay more attention to the development and use of educational games for language learning. Therefore, as a first contribution, this study focuses on developing a collaborative educational game for English vocabulary learning. In this context, Chiu, Kao, and Reynolds [15] highlighted the importance for further research on English learning using games beyond drill and practice genres.

Additionally, despite that educational games, including DGBLL, are motivating and interactive, they are black boxes [7], [8]. This means that teachers will not have the possibility to see how their students are learning (e.g., what they mastered and what not). To change that, researchers have thought of making use of the generated big data from the student's interaction with educational games by analyzing them to understand the learning process. The analysis of learning data is often referred to as LA, which is defined as "the measurement, collection, analysis and reporting of data about students and their context, for purposes of understanding and optimizing learning and the environments in which it occurs" [16]. Hauge et al. [17] mentioned that the provided feedback based on LA can help students do better in an educational game. Reinders [18] further mentioned that LA in language learning can help teachers monitor their students whether there are learning individually or collaboratively, hence provide early interventions and support accordingly. For instance, Youngs, Moss-Horwitz, and Snyder [19] applied LA for online French learning (not collaborative) and argued that one of the main purposes of LA is to provide teachers insights on student learning and highlight where and when they need to step in to monitor students. Consequently, the students had better learning performances compared to the other group of students (control group). Additionally, several studies showed that analyzing students' online behaviors could help in assessing their language learning performance and obstacles, hence those students who need more learning support could be identified [20], [21].

However, Gelan et al. [22] pointed out that little attention has been paid on the use of LA in language learning. Similarly, Thomas, Reinders and Gelan [9] stated that despite the promise of LA, its application in language teaching and learning has thus far been minimal. Hung, Yang, Hwang, Chu and Wang [23], in their literature review about GBLL, reported that most studies used traditional instruments instead of LA to evaluate the language learning process within games, namely perception questionnaires, learning tests and interviews. Nonetheless, in some studies studying LA in games [24]; [25], no study reported the use of LA in GBLL. This highlights the need for more practical investigations about the potential uses of LA in GBLL. Therefore, this study develops a smart collaborative educational game for teaching English, which incorporates LA to provide learning support for teachers to monitor their students while learning. Table I presents a comparison between various educational games for language learning that were developed in the literature, and the educational game reported in this study (last row of Table I). As the information shown in Table I, it can be seen that a lack of attention has been paid to develop collaborative educational games for English learning with LA support; this is the main contribution of this study.

TABLE I. Comparative Table of Educational Games in the Literature and the Developed Game in This Study

| Educational game | Taught Language | Learning type | Incorporate learning analytics |
|---|---|---|---|
| Hung, young and Lin [26] | English | Collaboratively | No |
| Hasegawa [27] | English | Individually | Yes |
| Wichadee and Pattanapichet [28] | English | Individually | Yes |
| Gamlo [29] | English | Individually | No |
| Bahari [30] | English | Collaboratively | No |
| Jungle animals | English | Collaboratively | Yes |

To summarize, this study contributes to extend the literature by developing a smart collaborative educational game, namely "Jungle animals game" for teaching English to primary school students. This game applied different collaborative strategies to enhance learning English vocabulary. It also incorporated LA approach based on K-mean algorithm to generate automatic dashboards for teachers to monitor their students and provide real-time interventions for them, as well as for students to keep track of their learning progress. K-means was chosen because of the simple implementation, speed of convergence and adaptability to sparse data [31].

## III. Jungle Animals Game

### A. Collaborative Learning

*Jungle animals* is an adventure 2D – multiplayer game that aims to teach English vocabulary, specifically animal names, as well as the spelling of each word for primary school students. The motivation behind choosing this topic (animal names) is that, most students are already familiar with the names of animals in their native language. Additionally, animals are a popular topic in early English curriculums [32]. In a face-to-face game setting, each student can see the other team members to ensure the individual accountability and increase collaboration. The story of the game is that a plane fell into an unknown forest. The survivors (students) must work together to find a computer with the GPS in it to escape. In the first level, the students have to find the password so they can use the computer. To do so, they have to collect letters and sounds which will be used for the password. During the second level, the students navigate to the nearest city where they will meet an old man to get instructions on how to make a city map by collecting pictures and words in the city, so they can find the right way to the airport to go home.

The students can be in different locations or in a shared area (forest in the proposed game) to increase interaction and collaboration among them [33]. Specifically, the shared "objects", such as letters, sounds and animals, can be found by different team members and they have then to exchange acquired information through social interaction. For instance, as shown in Fig. 1, a student in the first level cannot find the password without his/her team members' help because letters and sounds are divided among them. Therefore, they have to communicate via the chat box and work together to insert the password. Table II presents the different mechanisms used in the developed educational game in this study and how they promote collaboration.

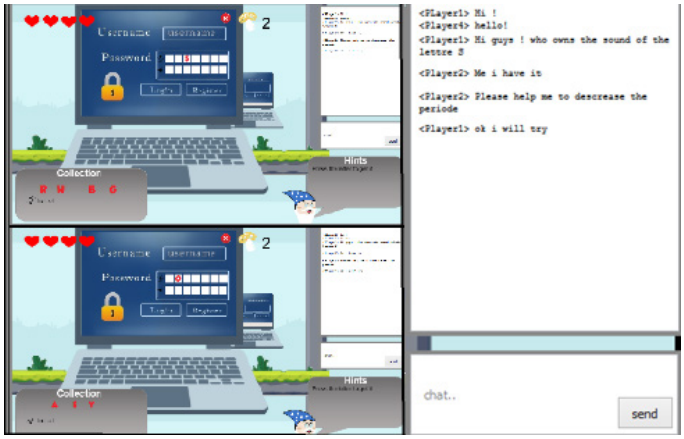



Fig. 1. An example of "insert password" collaborative task using chat box.

Furthermore, the game applies the receptive vocabulary knowledge based on the Nation's taxonomy [34]. It also supports the main parts of receptive knowledge of words, namely form and meaning. The receptive knowledge of word involves being able to recognize the form of the word when it is heard or met while reading. At the beginning of the game, different animal names are presented with their pictures. Students cannot move before hearing each word and seeing its spelling. The game also includes activities that allow students to exercise the spelling of the word in the second level, as shown in Fig. 2. For example, to collect pictures of animals, they have to work together to complete the missing letters.

TABLE II. The Implemented Collaborative Mechanisms in the Game

| **Principles of collaboration** | **Game mechanisms** | **Jungle animals Game** | |
|---|---|---|---|
| | | **Level 1** | **Level 2** |
| Reciprocity | • Collaborative task<br>• Encrypted information | **Members do the same effort/activities**:<br>• Each one is responsible on a part of the word and must put it in the correct place.<br>• Each student has parts of the password (letters or sounds). A student cannot put the password without the help of his/her group.<br><br>**Exchanging information:**<br>• Each student has a unique information and he/she must share it with his/her team members to find the password.<br>• The students can use the chat box to share information and communicate together. | **Members have the same effort/activities:**<br>• To collect the animal pictures, each student must write the missing letters.<br>• To collect part of the words, the students must complete the missing parts.<br><br>**Exchanging information:**<br>• The students can exchange information through the chat box. |
| Common goal | Collaborative task | Find the password using indices and write it with letters and sounds that are collected together. | Connect each animal with its name to complete the city map with the collected parts and pictures. |
| Trust and mutual respect | Collaborative task | • All roles are equally important.<br>• Without the participation of all students, the password cannot be written.<br>• Some letters are locked and the students need his/her team members to get it. | • Roles are exchanged between students. Thus, they all have the same value and importance.<br>• Without the participation of all the students, the city map cannot be completed. |
| Cohesiveness | • Shared space<br>• Shared object | Shared interface, shared letters, shared sounds. | Shared interface, shared pictures, shared words. |
| Benevolence and commitment | • Shared space<br>• Shared object | Ensuring the cohesiveness enhances individual benevolence and commitment. | Ensuring the cohesiveness enhances individual Benevolence and commitment. |

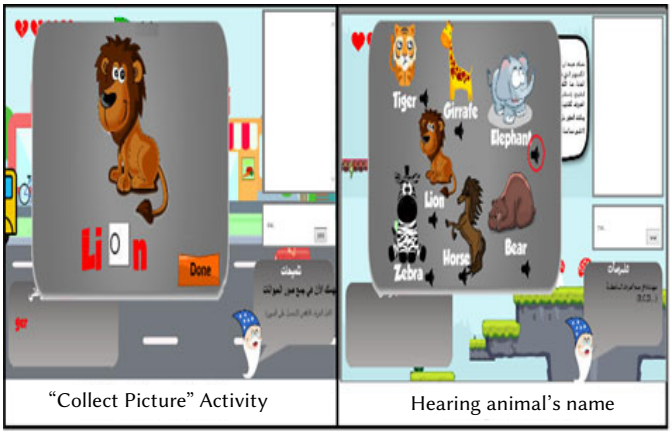



Fig. 2. Learning the form of words by filling the missing letters.

In addition, recognizing different portions of the word are also very important in the form part. Therefore, the game provides activities that develop recognizing portions of the words. Specifically, each student has a portion of the word, and they must collaborate together to combine all the pieces together before they can move to the next level. Regarding the meaning part, the knowledge of a word involves knowing its meaning. Therefore, the game adopts repetition and memory pedagogical strategy, proposed by Schmitt & McCarthy [35], to elicit word meaning. Memory strategy refers to relating the word with student's knowledge using images. In this context, students will encounter the same words and pictures in a repetitive cycle during different times at several activities.

The smart collaborative educational game aims to provide learning materials in a collaborative and fun way using tight coupling between text, speech and images to make the students learn how to pronounce each name and remember it when they see the picture of an animal. From the pedagogical perspective, it is difficult for teachers to keep up with each group and see how each student is behaving with his/her team members, as well as individual student's learning obstacles when it comes to learning the names of animals. For instance, a student might remember the outlook of an animal, but he/she still cannot spell the correct name of the animal. Also, a student might still have problems with memorizing the names of some animals. All these kinds of questions can be easily solved by providing dashboards to teachers and students through learning analytics (LA) approach.

### *B. Design of Learning Analytics*

Link and Li [36] highlighted several language learning interaction data that should be collected for LA based different theoretical approaches (e.g., interactionist, complexity, etc.). This study has relied specifically on "Skill acquisition theory" and "Interactionist theory", thus it analyzed performance data and communication activity data respectively (see Table III). Specifically, these traces can help teachers discover students' learning obstacles from three aspects, the individual aspect (the learning performance of each student), the group aspect (the learning performance of each group and the collaboration patterns among the members of each group) and the class aspect (the learning performance of the whole classroom. Consequently, each teacher will have the detailed information about how students learn from these three different aspects (individual, group and class).

During the process of game playing, time spent in different learning activities, wrong answers and retrying times were collected for each student and each group. The group communication patterns were also collected in order to evaluate the collaborative process during the game. Finally, the collected traces were automatically saved in an online database using PHP scripts. It should be noted that the smart collaborative educational game supports hundreds of students and vocabularies (animal names), and the limited number of students during this pilot experiment (see the next section) does not affect the technical reliability of the developed game.

TABLE III. The Collected Learning Traces

| Learning traces | Description |
|---|---|
| Time of solving activities | The total time that the student spends in each activity. |
| Number of wrong answers | Number of wrong answers made by the student before he/she finds the correct one during each activity. |
| Difficult activities | Activities that the student did not answer during the game. In addition, activities that student did not come out a solution until the game provides hints or answers. |
| Group time | The total time a group spent on achieving the common goal. |
| Group wrong answers | Number of wrong answers made by group members until they achieve the common goal of the activity. |
| Group communication | The number of discussing messages among the group members in the chat box. |

After collecting learning data, data mining and visualization techniques were applied to provide a detailed learning dashboard for teachers, to help teachers monitor their class and provide the needed interventions for each group or student accordingly. Gross, Stary and Totter [37] recommended that visualization tools for online learning should provide both group awareness and individual objective self-awareness. Group awareness presents information about group activities, collaborations and status [38], while individual objective self-awareness presents information on the process of taking oneself as the focus of one's behaviors and achievements [39].

An automatic game dashboard was created to display the current achievements of each individual student and group at any point of time. This dashboard shows how the students are progressing in the game. Specifically, it shows the number of completed activities by the students, as well as the number of correct and wrong answers (see Fig. 3). For example, most of the students did not answer correctly in the first learning activity (the yellow portion) as shown in Fig. 3, therefore the teacher should not move to the second learning activity, instead the teacher should help students answer the first learning activity by providing more explanation on the classroom blackboard. Additionally, the game dashboard shows the number of wrong answers given in each learning activity and the time spent in this activity. These features can provide teachers to have a global view about the learning difficulties of their students and help them accordingly. Similarly, the students can also access to the dashboards from their "student interface" to understand their learning performance and problems. For instance, students can see their own learning weakness and try to overcome it.

Furthermore, based on the group communication log data from the chat box of the game, the dashboard presents the interaction frequency in each group while collaborating. The communication frequency is also presented for each student while collaboratively solving different learning activities. Consequently, teachers can know if a group is experiencing certain collaboration problems and can check on them to provide needed help. Teachers can also know the communication frequency of each student during a specific period (e.g., the first level of the game) or during a specific activity (e.g., while solving the first learning activity), and understand how each student is involving in the learning process (as active or passive actor).

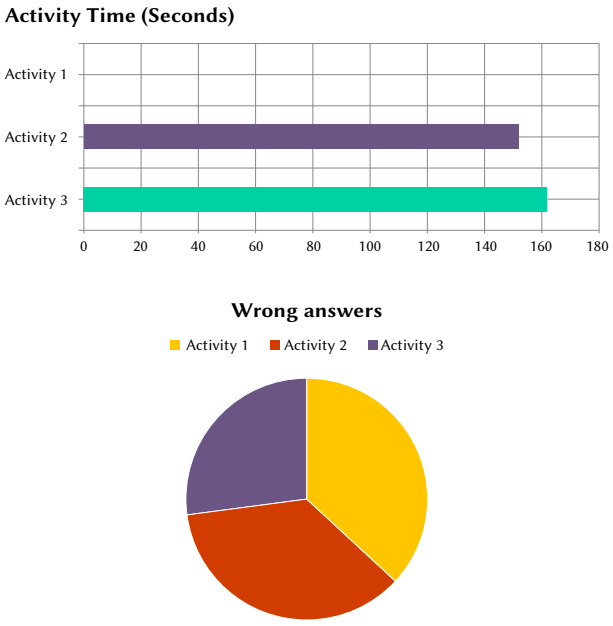



Fig. 3. Examples of learning dashboards generated by the smart collaborative educational game.

Finally, the dashboard automatically generates student clustering using K-means algorithm based on those with high, medium and low performance (see Fig. 4). This information can help teachers to provide required interventions to students according to different learning performance. The K-means algorithm divides the objects into k clusters, and iterates through the division-process as long as the distance between all objects and the center or mean of the clusters can be reduced. A characteristic of this algorithm is that the number k of clusters has to be fixed. In the game, the K value is fixed to 3 (similar to the above three groups) and two student features are used as inputs, namely number of wrong answers and time solving the activities. Fig. 4 shows the graphic representation of the three clusters with respect to the means of the two features. Since the first cluster has the highest mean of wrong answers and of time spent on activities, it is labeled "low performance". The second cluster is "high performance" and the final cluster is "medium performance".

| Name | Wrong answers | Time (Seconds) |
|---|---|---|
| Mo[redacted] | 21 | 123 |
| Ha[redacted] | 9 | 177 |
| Am[redacted] | 15 | 60 |
| Sa[redacted] | 12 | 102 |
| No[redacted] | 12 | 120 |
| Ra[redacted] | 12 | 132 |
| Do[redacted] | 15 | 123 |

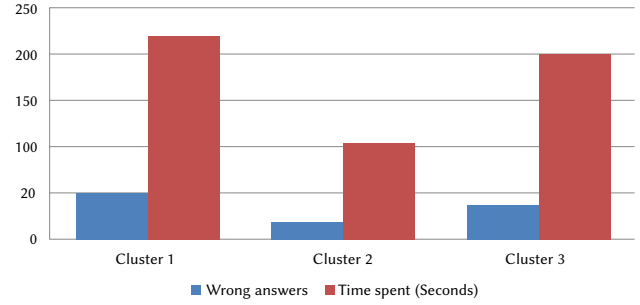



Fig. 4. Examples of the generated clusters using K-means algorithm by the smart collaborative educational game.

## IV. Method

Unlike educational games that incorporate LA, educational games without LA are black boxes and teachers cannot see how the learning process is occurring. Hence, they cannot provide real-time learning support to help students achieve better learning outcomes. Additionally, students cannot receive any instant feedback about their learning progress and performance. Based on this, this study aimed to validate the following two hypotheses.

**H1**: Students learning with the developed smart collaborative educational game (with LA) have significantly higher learning performance than students learning with the non-smart version of the game (without LA).

**H2**: Students learning with the developed smart collaborative educational game have (with LA) significantly higher learning motivation during the learning process than students learning with the non-smart version of the game (without LA).

### *A. Participants*

A pilot experiment, during the academic year 2018-2019, was conducted to validate the two hypotheses at a public primary school after receiving the approval from school review board (including parents' consent forms). Thirty-one sixth-grade primary school students participated in this study where 70% of them were boys and 30% were girls. The average age of the students was 12 years old. The students were randomly divided into two groups, namely experimental and control groups.

### *B. Experimental Procedure*

The teacher started by introducing the game to the students in both groups (fifteen minutes for each group). After that, the students in each group (experimental and control) took forty-five minutes to answer a pre-test and a pre-motivation questionnaire to assess their prior-knowledge and motivation related to English vocabulary. The students in the experimental group then used the smart collaborative educational game (with LA) to learn English, while the students in the control group used the non-smart version of the educational game (without LA). The learning process of both groups was in different time slots (morning and afternoon) in order for the same teacher to facilitate both learning processes. The learning process was for three hours for each group (control and experimental). The game included twenty animal names, where ten of them are herbivore and the other ten are carnivore. The students were divided into different teams with four members in each team; two members with high English vocabulary achievements and two members with low English vocabulary achievements (based on their English test results from the previous academic year). Finally, after the learning process was ended, the students in both groups completed a post-test and a post-motivation questionnaire.

### *C. Instruments and Data Collection*

Both qualitative and quantitative data from both the students and the teacher are collected using the following three instruments. The main idea is that the results from qualitative analysis should further support and explain the quantitative results.

- *Pre and post-test*: It was designed by experienced teachers who had taught English courses in primary school for the past fifteen years. This test contains three different items and aims to measure each student's learning performance regarding animal names learned during the game. For instance, in the first item, students were requested to fill the missing letters of a particular given name of an animal. In another item, students were requested to link using arrows the animal picture with its correct name, among several

provided names. The students took between 25 and 30 minutes to finish this test. It should be noted that the pre and post-tests are the same and 10 is the highest grade that a student can obtain.

- *Pre and post-motivation questionnaire*: The motivation questionnaire was adapted from Wigfield and Guthrie [40]. It aims to measure the motivation level of students during the learning process using the game. It consists of nine items on a four-point scale (1 strongly disagree; 2 disagree; 3 agree; and, 4 strongly agree). The Cronbach's alpha of the questionnaire was calculated and it was equal to 0.83. This implied that it was reliable since Cronbach's alpha value was greater than 0.7 [41]. It should be noted that the pre and post-motivation questionnaires are the same.
- *Interview*: A semi structured interview was conducted with the teacher to collect his feedback about using the smart collaborative educational game and the non-smart version of it for teaching English vocabulary. The interview took 30 minutes and it was recorded in order to be analyzed and draw conclusions. The coding process was done by two coders, and in case of disagreement, the two researchers resolved it through discussion. Specifically, four codes were used for the qualitative analysis of interviews, namely: (1) Learning obstacle: Use this code when the teacher is talking about how using the smart and non-smart versions of the educational game helped him in identifying the learning obstacle (difficulties, wrong answers, etc.) of students; (2) Timely intervention: Use this code when the teacher is talking about how using the smart and non-smart versions of the educational game helped him in providing immediate or effective learning interventions; (3) Communication: Use this code when the teacher is talking about communication and interaction between students while using the smart and non-smart versions of the educational game; and, (4) Reflection: Use this code when the teacher is talking about students' self-reflection while using the smart and non-smart versions of the educational game.
- *Observation*: During the learning processes (using the smart collaborative educational game and the non-smart version of it), two observers were in the classrooms to observe the effects of the two games on the learning behaviors of students. The two chosen observers are teachers with more than twenty years teaching experience, and they did not have any previous relationship with the students. The coding process was done by two coders (same coders who coded the interview), and in case of disagreement, the two coders resolved it through discussion. Specifically, the coders mainly focused on two aspects which can affect the learning motivation, namely interactivity [37], [38] and exhibiting excitement and fun [39]. Specifically, two codes were used for the qualitative analysis of observations, namely: (1) Interaction: Use this code for all occurrences that illustrate teacher-student, student-student or student-game interactions while using the smart and non-smart version of the collaborative educational game; and, (2) Excitement/Fun: Use this code for all occurrences that illustrate students are exhibiting excitement or fun while using the smart and non-smart version of the collaborative educational game.

## V. Results

### A. Impacts on Learning Performance (Hypothesis 1)

The pre-test scores of both groups (control and experimental) were analyzed using the two sample t-test which was reported as an effective statistical method to deal with limited sample size [40], as shown in Table IV. The obtained results showed that there was no significant difference in the pre-test performance of both groups since the p value was equal to .066 and greater than .005. To conclude, there was no significant difference in the prior-knowledge of English vocabulary between the control and experimental groups before the beginning of the learning process.

TABLE IV. Two-sample T-test Results of the Pre-tests Analysis

| Pair 1 | Mean | SD | t | df | Sig |
|---|---|---|---|---|---|
| Pre_testl & Pre_test2 | 1.6 | .66 | -3.1 | 14 | .066 |

After the learning process, the post-test scores were analyzed using the two-sample t-test, as shown in Table V. The obtained results showed that there was a significant difference in the post-test performance of both groups since the p value was equal to .001 and less than .05. Specifically, the experimental group achieved higher scores in the post-tests of English vocabulary than the control group.

TABLE V. Two-sample T-test Results of the Post-tests Analysis

| Pair 2 | Mean | SD | t | df | Sig |
|---|---|---|---|---|---|
| Post_test1 & Post_test2 | 6.22 | 3.25 | 4.37 | 14 | .001 |

To understand how the smart collaborative educational game helped the experimental group achieving a better learning performance, the teacher was interviewed and the given answers were qualitatively analyzed. The distribution rate of each coding item is presented in Fig. 5. Specifically, it can be seen from these bar chart that the smart collaborative educational game was more helpful for the teacher than the non-smart version of it. To better understand the obtained results of each coding distribution, the interview answers were analyzed and discussed as follows:

- *Learning weakness*: The teacher reported that the provided dashboards in the smart version of the educational game helped him to identify the learning weakness of the students (individually or in groups). However, this was not very easy when he used the non-smart version of the game since he had to go through every team and keep an eye on their computer screens to see how they are performing, as no feedback was given to him (i.e., the game was a black box). For instance, the teacher mentioned that, from the provided LA dashboard, he could easily see that some students still cannot spell correctly "giraffe" and "elephant". He also mentioned that the smart collaborative educational game helped him automatically identify students with different learning performances (low, medium and high). For instance, the teacher mentioned that he can easily see that the student <name withheld> was struggling to solve the first activity compared to his team members.
- *Communication*: The teacher mentioned that both educational games (the smart and the non-smart version of it) enhanced the communication level between the students as they both support the collaborative learning strategy. This was further reflected in the "communication" bar chart in Fig. 5, as no huge difference was seen. However, the teacher mentioned that the provide LA dashboards within the smart collaborative educational game made the students more interactive compared to the students who used the non-smart version of it. For instance, every time the students see their team performance, through the LA dashboards, compared to the other teams, they start discussing their learning-playing strategies to increase their winning chances. The teacher further mentioned that the students sometimes leave their seats and go to their peers to talk to them, instead of using the chat box. This was encouraging and helpful in a way that the students were motivated to learn from each other.

- *Timely intervention*: As discussed in the first coding scheme, unlike the non-smart version of the educational game, the smart educational game provided detailed information using dashboards to the teacher about the learning weakness of his students. Therefore, he provided timely interventions accordingly. For instance, when he noticed that some students still cannot spell correctly "giraffe" and "elephant", he helped them write it down on the board couple of times to memorize it. Also, he instantly provided help to the student <name withheld> in order to correctly finish the first activity. Furthermore, the teacher mentioned that every time he sees that the communication frequency of some groups is low, he goes there to encourage them to communicate together. Finally, the teacher mentioned that the provided dashboards helped him assess his class performance and identify their weakness, hence easily identify the supplemental learning materials that he needed to suggest.
- *Reflection*: The teacher mentioned that both educational games (the smart and the non-smart version of it) through the collaborative strategy helped students to have self-reflection about their actions and achievements while communicating with their team members via the chat box about their learning-playing strategies to win. The teacher, however, mentioned that the LA dashboards specifically, within the smart collaborative educational game, further emphasized self-reflection by summarizing the learning progress of each student in simple dashboards. Consequently, it is seen that several students refer to the dashboard to see their learning weakness and then start consulting their peers via the chat box for help.

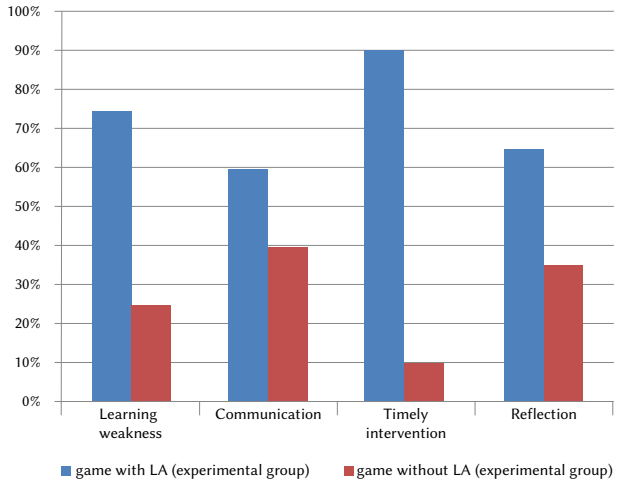



Fig. 5. Distribution of the four interview features based on the used version of the game.

## B. Impacts on Motivation Level (Hypothesis 2)

Similar to the first analysis, the pre-motivation questionnaire scores of both groups were analyzed using two sample t-test as shown in Table VI. The obtained results showed no significant difference in the motivation levels between the experimental and control groups towards learning English vocabulary before the experiment. Particularly, the p value was equal to .41 and greater than .05.

TABLE VI. Two-sample T-test Results of the Pre-motivation Questionnaire Analysis

| Pair 1 | Mean | SD | t | df | Sig |
|---|---|---|---|---|---|
| Pre_quest1 & Pre_quest2 | 1.21 | .77 | -4.72 | 13 | .41 |

After the learning process, the post-motivation questionnaire scores were analyzed as well using the two-sample t-test, as shown in Table VII. The obtained results showed that there was a significant difference in the post-motivation questionnaire scores of the two groups since the p value was equal to .01 and less than .05. Specifically, the experimental group had a higher motivation level towards learning English vocabulary than the control group.

TABLE VII. Two-sample T-test Results of the Post-motivation Questionnaire Analysis

| Pair 2 | Mean | SD | t | df | Sig |
|---|---|---|---|---|---|
| Post_quest1 & Post_quest2 | 3.81 | 1.81 | -2,14 | 13 | .01 |

To understand how the smart collaborative educational game helped the experimental group achieving a higher motivation level, the observations of both learning processes (using the smart collaborative educational game and the non-smart version of it) were qualitatively analyzed. The distribution rate of each coding item is presented in Fig. 6. Specifically, it can be seen from these bar chart that the smart collaborative educational game made students more interactive and exhibit high level of fun and excitement than the students who used the non-smart version of it. Consequently, these students had higher motivation level. To better understand the obtained results of each coding distribution, the collected observations were analyzed and discussed as follows:

- *Interaction*: It is evidenced that the LA dashboard provided by the smart collaborative educational game made the students very active and engaged. This is seen when they always refer to this dashboard to start discussing strategies to win or helping each other to increase their chances of winning. This created a motivating atmosphere while learning. When using the non-smart version of the game, interaction was relatively low among students due to the absence of dashboards, where they discussed only the learning-playing process. However, it was seen that some students asked directly their friends about their performance and some learning conversation happened as a result.
- *Excitement/Fun*: It is evidenced that the provided learning dashboards by the smart collaborative educational game made the students very excited. Specifically, it was frequently seen that the students in each team expressed excitement when they referred to the dashboards and saw that they are wining and. However, this was not the case in the non-smart version of the game. Particularly, the students expressed high excitement level only at the beginning of the learning process (during the first 15 or 20 minutes) since using the game was fun for them.

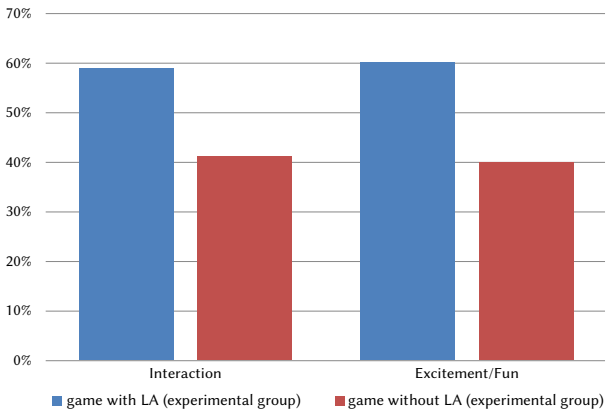



Fig. 6. Distribution of the two observation features based on the used version of the game.

## VI. Conclusions, Discussions and Implications

This study developed and validated a smart collaborative educational game incorporated with LA to teach English vocabulary. The first obtained results showed that the students who learned English vocabulary using the smart collaborative educational game achieved a higher learning performance than students who used the non-smart version of the game. This can be explained by the automatically generated dashboards by the smart collaborative educational game for teachers to get real-time information about their students' learning situations and provide the needed interventions in a timely manner. From the pedagogical perspective, Reinders [18] found that LA can help teachers monitor their students whether they are learning individually or collaboratively, hence provide early interventions and support accordingly. Additionally, the teacher during the conducted interview revealed that displaying team achievements using LA dashboards to the students could also help them perform better. In this context, several researchers mentioned that providing learning achievements information of individuals and groups in collaborative environments could enhance online participation and learning performances [37].

The second obtained results showed that students who used the smart collaborative educational game had a higher motivation level than the students who used the non-smart version of the educational game. This could be attributed to the smart collaborative educational game can facilitate self-reflection via the provided dashboards (as reported in the interview results), this has affected positively the students' learning motivation and outcomes. In this context, several research studies showed that supporting self-reflection can enhance students' learning motivation [46], [47]. Particularly, the information displayed on the dashboard provided by the smart collaborative educational game to the students about their learning progress made them more excited and encouraged them to do better, hence they were very motivated. In this context, Wang [48] stated that an educational game can motivate students while learning, but their motivation level will start decreasing once they get familiar with the game. Therefore, incorporating motivational strategies to encourage continuous play is crucial [49]. An effective motivational technique in education is to highlight a student's accomplishments [50], thus LA dashboards that visualize a student's improvement could be motivating. Additionally, it had seen that the smart collaborative educational game, through the provided dashboards made the students more interactive by collaborating together to win than the students who used the non-smart version of the game. Similarly, several studies also showed that providing interactive learning process can positively affect the students' learning motivation [42]; [43].

The findings of this research could enhance the educational technology field by presenting a new learning tool (smart collaborative educational game) that can collaboratively help in learning English vocabulary. Specifically, this study presented examples of implementing game mechanics and scenarios that other researchers and practitioners could apply in their respective educational game contexts to fulfill different collaborative learning strategies. For instance, to fulfill memory strategy, students will encounter, during the game, the same words and pictures in a repetitive cycle during different times at several activities. This will elicit their memory and help them recall the learned knowledge. Some suggestions to the designers and teachers learned from this study are: (1) focus not only on the learning perspective (performance, weakness and progress), but also on the social perspective as well (communication between peers/teachers); (2) provide feedback during the learning process about both individual and team achievements; and, (3) provide simple interfaces (dashboards) without detailed information (using pie chart, histograms, etc.) to help teachers/students easily identify important information and make use of it.

It should be noted that this study has several limitations that should be acknowledged and further investigated. For instance, the sample size of the experiment was limited, due to the experiment context (public school). Also, the learning process of each group (control and experimental) was only for three hours. However, despite these limitations, this study presented insights, including practical examples and recommendations for applying both collaborative learning as well as learning analytics in DGBLL. Future research work could focus on making the designed game smarter by providing automatic learning support and interventions based on different learning scenarios and conditions. For instance, when a team is having low communication frequency, the game will start providing encouragements for students to make them more active and share ideas together. In addition, future directions could focus on designing a mobile version of this game, as language learning games are gaining an increasing attention on mobile devices [51].

## References


[1] M. Arnold, S. C. Greenville, & R. Doe, "Second Language Acquisition Video Game." Accessed: Oct. 15, 2019. [Online]. Available: https://cse.sc.edu/files/Matthew%20and%20Renaldo.pdf.

[2] J. F. F. Flores, "Using gamification to enhance second language learning," *In Digital Education Review*, vol. 27, no. 21, pp. 32-54, 2015.

[3] M. Amoia, T. Brétaudière, A. Denis, C. Gardent, & L. Perez-Beltrachini, "A serious game for second language acquisition in a virtual environment," *Journal on Systemics, Cybernetics and Informatics (JSCI), International Institute of Informatics and Systemics*, vol. 10, no. 1, pp. 24-34, 2012.

[4] C. Y. Hung, J. C. Y. Sun, & P. T. Yu, "The benefits of a challenge: student motivation and flow experience in tablet-PC-game-based learning." *Interactive Learning Environments*, vol. 23, no. 2, pp. 172-190, 2015, doi: 10.1080/10494820.2014.997248.

[5] M. Filsecker, & D. T. Hickey, "A multilevel analysis of the effects of external rewards on elementary students' motivation, engagement and learning in an educational game," *Computers & Education*, vol. 75, pp. 136-148, 2014, doi: 10.1016/j.compedu.2014.02.008.

[6] A. Hawlitschek, & S. Joeckel, S. "Increasing the effectiveness of digital educational games: The effects of a learning instruction on students' learning, motivation and cognitive load," *Computers in Human Behavior*, vol. 72, pp. 79-86, 2017, doi: 10.1016/j.chb.2017.01.040.

[7] C. Alonso-Fernandez, A. Calvo, M. Freire, I. Martinez-Ortiz, & B. Fernandez-Manjon, "Systematizing game learning analytics for serious games," in *2017 IEEE Global Engineering Education Conference (EDUCON)*, Athens, Greece, 2017, pp. 1111-1118.

[8] A. Tlili, & M. Chang, M. "Data Analytics Approaches in Educational Games and Gamification Systems: Summary, Challenges, and Future Insights," in *Data Analytics Approaches in Educational Games and Gamification Systems*, Springer, Singapore, 2019, pp. 249-255.

[9] M. Thomas, H. Reinders, & A. Gelan, "Learning analytics in online language learning: Challenges and future directions." in *Faces of English Education*, Oxfordshire, England, UK, Routledge, 2017, pp. 197-21.

[10] D. Matukhin, & D. Bolgova, "Learner-centered Approach in Teaching Foreign Language: Psychological and Pedagogical Conditions," *Procedia-Social and Behavioral Sciences*, vol. 206, pp.148-155, 2015.

[11] B. Hart, & T. R. Risley, *Meaningful differences in the everyday experience of young American children*, Paul H Brookes Publishing, 1995.

[12] D. Siemon, F. Becker, L. Eckardt, & S. Robra-Bissantz, "One for all and all for one-towards a framework for collaboration support systems," *Education and Information Technologies*, vol. 24, no. 2, pp. 1837-1861, 2019, doi: 10.1007/s10639-017-9651-9.

[13] C. Surkamp, & B. Viebrock, B., *Teaching English as a Foreign Language: An Introduction*, Stuttgart, Germany: JB Metzler: Springer, 2018.

[14] H. T. Hung, J. L. Chang, & H. C. Yeh, "A review of trends in digital game-based language learning research," in *16th International Conference on Advanced Learning Technologies (ICALT)*, Austin, TX, United States, 2016, pp. 508-512.

[15] Y. H. Chiu, C. W. Kao, & B. L. Reynolds, "The relative effectiveness of digital game-based learning types in English as a foreign language

setting: A meta-analysis," *British Journal of Educational Technology,* vol. 43, no. 4, pp. 104-107, 2012.

[16] G. Siemens, & P. Long, "Penetrating the fog: Analytics in learning and education," *EDUCAUSE review,* vol. 46, no. 5, p. 30, 2011.

[17] J. B. Hauge, R. Berta, G. Fiucci, B. F. Manjón, C. Padrón-Nápoles, W. Westra, & R. Nadolski, "Implications of learning analytics for serious game design,"in *14th international conference on advanced learning technologies,* Athens, Greece, 2014, pp. 230-232.

[18] H. Reinders, "Learning Analytics for Language Learning and Teaching." *JALT CALL Journal*, vol. 14, no. 1, pp.77-86, 2018.

[19] B. Youngs, S. Moss-Horwitz, & E. Synder, "Educational data mining for elementary French on-line: A descriptive study," in *E. Dixon & M. Thomas (Eds.), Researching language learner interactions online: From social media to MOOCs*, Texas: CALICO, 2015, pp. 347-368.

[20] E. Martín-Monje, M. D. Castrillo, & J. Mañana-Rodríguez, "Understanding online interaction in language MOOCs through learning analytics," *Computer Assisted Language Learning*, vol. 31, no. 3, pp. 251-272, 2018.

[21] F. Rubio, J. M. Thomas, & Q. Li, "The role of teaching presence and student participation in Spanish blended courses. *Computer Assisted Language Learning*, vol. 31, no. 3, pp. 226-250, 2018.

[22] A. Gelan, G. Fastré, M. Verjans, N. Martin, G. Janssenswillen, M. Creemers,., ... & M. Thomas, "Affordances and limitations of learning analytics for computer-assisted language learning: a case study of the VITAL project," *Computer Assisted Language Learning*, vol. 31, no. 3, pp. 294-319, 2018.

[23] H. T. Hung, J. C. Yang, G. J. Hwang, H. C. Chu, & C. C. Wang, "A scoping review of research on digital game-based language learning," *Computers & Education*, vol. 126, pp. 89-104, 2018, doi: 10.1016/j.compedu.2018.07.001.

[24] C. Alonso-Fernández, A. Calvo-Morata, M. Freire, I. Martínez-Ortiz, & B. Fernández-Manjón, "Applications of data science to game learning analytics data: A systematic literature review," *Computers & Education*, vol. 141, 2019, doi: 10.1016/j.compedu.2019.103612.

[25] M. Freire, Á. Serrano-Laguna, B. Manero, I. Martínez-Ortiz, P. Moreno-Ger, & B. Fernández-Manjón, "Game learning analytics: learning analytics for serious games," Learning, design, and technology Springer Nature Switzerland AG, pp. 1-29, 2016, doi: 10.1007/978-3-319-17727-4_21-1

[26] H. C., Hung, S. S. C., Young, & C. P., Lin, "No student left behind: a collaborative and competitive game-based learning environment to reduce the achievement gap of EFL students in Taiwan." *Technology, Pedagogy and Education*, vol. 24, no. 1, pp. 35-49, 2015, doi: 10.1080/1475939X.2013.822412.

[27] Hasegawa, et al., "An English vocabulary learning support system for the learner's sustainable motivation," *Springer Plus*, 2015, doi: 10.1186/s40064-015-0792-2.

[28] S., Wichadee, & F., Pattanapichet, "Enhancement of performance and motivation through application of digital games in an English language class," *Teaching English with Technology*, vol.18, no.1, pp. 77-92, 2018.

[29] N., Gamlo, "The Impact of Mobile Game-Based Language Learning Apps on EFL Learners' Motivation." *English Language Teaching*, vol.12, no.4, pp.49-*56,* 2019, doi: 10.5539/elt.v12n4p49.

[30] *A., Bahari, "Game-based collaborative vocabulary learning in blended* and distance L2 learning." Open Learning: *The Journal of Open, Distance and e-Learning*, vol. 35, no. 3, pp.34-59, 2020, 10.1080/02680513.2020.1814229.

[31] A. C. Fabregas, B. D. Gerardo, & B. T. Tanguilig III, "Enhanced initial centroids for k-means algorithm," *Int. J. of Information Technology and Computer Science*," vol. 9, no. 1, pp.26-33, 2017, doi: 10.5815/ijitcs.2017.01.04.

[32] J. Sandberg, M. Maris, & K. De Geus, "Mobile English learning: An evidence-based study with fifth graders," *Computers & Education*, vol.57, no. 1, 2011, pp. 1334-1347, doi: 10.1016/j.compedu.2011.01.015.

[33] E. Szewkis, M. Nussbaum, T. Rosen, J. Abalos, F. Denardin, D. Caballero, .... & C. Alcoholado, "Collaboration within large groups in the classroom," *International Journal of Computer-Supported Collaborative Learning*, vol. 6, no. 4,, pp. 561-575, doi: 2011, 10.1007/s11412-011-9123-y.

[34] I. S. Nation, *Learning vocabulary in another language*, Cambridge University Press, 2011.

[35] N. Schmitt, & M. McCarthy, M., *Vocabulary: Description, acquisition and pedagogy*, Cambridge university press, 1997.

[36] S. Link, & Z. Li, "Understanding online interaction through learning analytics: Defining a theory-based research agenda," in *E. Dixon & M. Thomas (Eds.), Researching language learner interactions online: From social media to MOOCs*, San Marcos, TX: CALICO Monograph Series, 2015, pp. 369–385.

[37] T. Gross, C. Stary, & A. Totter, "User centered awareness in computer supported cooperative work systems: Structured embedding of findings from social sciences," *International Journal of Human Computer Interaction*, vol. 18, no.3, pp. 323-360, 2005 doi: 10.1207/s15327590ijhc1803_5.

[38] S. Greenberg, C. Gutwin, & A. Cockburn, "Awareness through fisheye views in relaxed WYSIWIS groupware" in Proceedings of Graphics interface, Toronto, Canada, 1996, pp.28-38.

[39] B. Mullen, & G. R. Goethals, *Theories of group behavior*, New York, USA: Springer Science & Business Media, 1987.

[40] A. Wigfield, & J. T. Guthrie, "Relations of children's motivation for reading to the amount and breadth or their reading," *Journal of educational psychology*, vol. 89. No. 3, pp. 420, 1997, doi: 10.1037/0022-0663.89.3.420.

[41] C. H. Yu, "An introduction to computing and interpreting Cronbach Coefficient Alpha in SAS," in *Proceedings of 26th SAS User Group International Conference,*. Cary, NC: SAS Institute Inc, 2001 pp. 1-6.

[42] S. Ekiz, & Z. Kulmetov, "The factors affecting learners' motivation in English language education," *Journal of Foreign Language Education and Technology*, vol. 1, no. 1, 2016.

[43] E. L. Snow, G. T. Jackson, L. K. Varner, & D. S. McNamara, "The impact of system interactions on motivation and performance in a game-based learning environment," in *International Conference on Human-Computer Interaction*, Berlin, Germany, 2013, pp. 103-107.

[44] A. Tlili, F. Essalmi, & M. Jemni, "Improving learning computer architecture through an educational mobile game." *Smart Learning Environments*, vol. 3, no. 1, 2016, pp. 7.

[45] J. Sauro, "Best Practices For Using Statistics On Small Sample Sizes." Accesse: Feb. 07, 2020. [Online]. Available: https://measuringu.com/small-n/

[46] B. A. Greene, R. B. Miller, H. M. Crowson, B. L. Duke, & K. L. Akey, "Predicting high school students' cognitive engagement and achievement: Contributions of classroom perceptions and motivation," *Contemporary Educational Psychology*, vol. 29, pp. 462–482, 2004, doi: 10.1016/j.cedpsych.2004.01.006.

[47] H. W. Stevenson, C. Chen, & S. Lee, "Motivation and achievement of gifted children in East Asia and the United States," *Journal for the Education of the Gifted*, vol. 16, pp. 223–250, 1993.

[48] A. I. Wang, "The wear out effect of a game-based student response system." *Computers & Education*, vol. 82, pp. 217-227, 2015, doi: 10.1177/016235329301600302.

[49] P. Wouters, C. Van Nimwegen, H. Van Oostendorp, & E. D. Van Der Spek, "A meta-analysis of the cognitive and motivational effects of serious games," *Journal of educational psychology*, vol. 105, no. 2, pp. 249, 2013, doi: 10.1037/a0031311.

[50] J. M. Keller, "Strategies for stimulating the motivation to learn," *Performance+ Instruction*, vol. 26, no. 8, pp. 1-7, 1987.

[51] E. Nuñez-Valdez, J. M. Cueva-Lovelle, C. P. G-Bustelo, G. Infante-Hernandez, O. Sanjuan-Martinez, "Gade4all: developing multi-platform videogames based on domain specific languages and model driven engineering." *International Journal of Interactive Multimedia And Artificial Intelligence*, vol. 2. No. 2., pp. 33-42, 2013, doi: 10.9781/ijimai.2013.224.

Ahmed Tlili

He is the Co-Director of the OER Lab at the Smart Learning Institute of Beijing Normal University (SLIBNU), China. He serves as the Associate Editor of the IEEE Bulletin of the Technical Committee on Learning Technology, and the Journal of e-Learning and Knowledge Society. He is also a Visiting Professor at UNIR-iTED, Spain, and an expert at the Arab League Educational, Cultural and Scientific Organization (ALECSO). Dr. Tlili has been awarded the IEEE TCLT Early Career Researcher Award in Learning Technologies for 2020. He is the Co-Chair of IEEE special interest group on "Artificial Intelligence and Smart Learning Environments" and APSCE's Special Interest Group on "Educational Gamification and Game-based Learning (EGG)". His research interests include, open education, game-based learning, educational psychology and artificial intelligence.

Sarra Hattab

She has a master degree in Intelligent information systems. Her research focuses on language learning using educational games.

Fathi Essalmi

He is the former Head of the Computer Science Department at Kairouan University, Tunisia. He is currently an Assistant Professor at the University of Jeddah, Saudi Arabia. He supervises master's students and co-supervises Ph.D. students in two fields: learner modeling based on computer games and federation of personalization efforts. He has several publications with international team appeared in journals with impact factor and ranked conferences. He is also a program committee member in several conferences and a reviewer in several journals.

Nian-Shing Chen

He is currently Chair Professor in the Department of Applied Foreign Languages at the National Yunlin University of Science and Technology, Taiwan. He has published over 400 academic papers in the international referred journals, conferences and book chapters. One of his papers published in Innovations in Education and Teaching International was awarded as the top cited article in 2010. He is the author of three books with one textbook entitled "e-Learning Theory & Practice". He has received the national outstanding research awards for three times from the National Science Council in 2008, 2011-2013 and the Ministry of Science and Technology in 2015-2017.

Ronghuai Huang

He is a Professor in Faculty of Education of Beijing Normal University (BNU). He has being engaged in the research on smart learning environment, artificial intelligence in education, educational technology as well as knowledge engineering. He received 'Chang Jiang Scholar' award in 2016, which is the highest academic award presented to an individual in higher education by the Ministry of Education of China. He serves as Co-Dean of Smart Learning Institute, Director of UNESCO International Rural Educational and Training Centre, and Director of China National Engineering Lab for Cyber learning Intelligent Technology. He is also the Editor-in-Chief of Springer's Journal of Smart Learning Environment and Journal of Computers in Education. Till now, he has finished over 100 projects, and published over 400 academic papers and 40 books.

Kinshuk

He is currently a Professor of Computer Science and the Dean of the College of Information at the University of North Texas, USA. He received the Ph.D degree in computer science from the University of De Montfort, England, in 1996. He held the NSERC/CNRL/ Xerox/McGra Hil Research Chair for Adaptivity and Personalization in Informatics, funded by the Federal government of Canada, Provincial government of Alberta, and by national and international industries. Areas of his research interests include learning analytics; learning technologies; mobile, ubiquitous and location aware learning systems; cognitive profiling; and, interactive technologies.

Maiga Chang

He is currently a Full Professor with the School of Computing Information and Systems, Athabasca University, Canada. He has given more than 105 talks and lectures in different conferences, universities, and events. He has participated in more than 310 international conferences and workshops as a Program Committee Member. He has (co-)authored more than 225 edited books, special issues, book chapters, journal and international conference papers. He is the Editor-in-Chief of the Educational Technology and Society, the Bulletin of Technical Committee on Learning Technology, and the International Journal of Distance Education Technologies. He was a Section Editor of the Education and Science, an Associate Editor of Transactions on Edutainment (Springer). He is an Advisory Board Member of the Journal of Computers and Applied Science Education. He is also the Chair of the IEEE Technical Committee of Learning Technology (IEEE TCLT), an Executive Committee Member of the Asia-Pacific Society for Computers in Education (APSCE), the Global Chinese Society for Computing in Education (GCSCE), and the Chinese Society for Inquiry Learning (CSIL). He is a Secretary and a Treasurer of the International Association of Smart Learning Environments (IASLE).

Daniel Burgos

He received a postgraduate in artificial intelligence & machine learning from MIT, and the Ph.D. degree in communication, the Dr.Ing. degree in computer science, the Ph.D. degree in education, the Ph.D. degree in anthropology, and the D.B.A. degree in business administration. He is currently as a Full Professor of technologies for education & communication and the Vice-Rector for International Research, the UNESCO Chair of eLearning, and the ICDE Chair of open educational resources with the Universidad Internacional de La Rioja. He is also the Director of the Research Institute for Innovation & Technology in Education (UNIR iTED). He has published over 150 scientific articles, 20 books, and 15 special issues on indexed journals. He has developed +55 European and Worldwide Research and Development projects. His research interests include adaptive, personalised and informal eLearning, learning analytics, open education and open science, eGames, and eLearning specifications.

# Your Teammate Just Sent You a New Message! The Effects of Using Telegram on Individual Acquisition of Teamwork Competence

Miguel Á. Conde[1]*, Francisco J. Rodríguez-Sedano[1], Ángel Hernández-García[2], Alexis Gutiérrez-Fernández[1], Ángel M. Guerrero-Higueras[1]

[1] Universidad de León, León (Spain)
[2] Universidad Politécnica de Madrid, Madrid (Spain)

Received 25 March 2021 | Accepted 26 April 2021 | Published 18 May 2021

## Abstract

Students' acquisition of teamwork competence has become a priority for educational institutions. The development of teamwork competence in education generally relies in project-based learning methodologies and challenges. The assessment of teamwork in project-based learning involves, among others, assessing students' participation and the interactions between team members. Project-based learning can easily be handled in small-size courses, but course management and teamwork assessment become a burdensome task for instructors as the size of the class increases. Additionally, when project-based learning happens in a virtual space, such as online learning, interactions occur in a less natural way. This study explores the use of instant messaging apps (more precisely, the use of Telegram) as team communication space in project-based learning, using a learning analytics tool to extract and analyze student interactions. Further, the study compares student interactions (e.g., number of messages exchanged) and individual teamwork competence acquisition between traditional asynchronous (e.g., LMS message boards) and synchronous instant messaging communication environments. The results show a preference of students for IM tools and increased participation in the course. However, the analysis does not find significant improvement in the acquisition of individual teamwork competence.

## Keywords

Instant Messaging, Learning Analytics, Students Interaction, Telegram, Teamwork.

DOI: 10.9781/ijimai.2021.05.007

## I. Introduction

PREPARING students to be successful workers or entrepreneurs is one of the main goals of educational institutions. For success to happen, students need to acquire competences that are demanded by the labor market to increase their employability and performance. However, this is easier said than done, as it requires that: 1) companies and educational institutions match learning curricula and business requirements; and 2) educational institutions use the adequate tools that facilitate competence acquisition and assessment [1]. This study addresses both issues. The first one, by exploring the application of a methodology to facilitate students the acquisition of Teamwork Competence (TWC) in educational contexts. The second, by exploring the application of a Learning Analytics tool to facilitate assessment of TWC.

The acquisition of TWC has multiple benefits for students' learning and development [2], [3] and is highly demanded by companies [1]. Different methodologies may facilitate its acquisition; most of them require that students work together in groups to develop a project or solve some problem or challenge. These methodologies share a common hurdle: while assessing the final result of a group (e.g., the project delivered) is easy and quite straightforward, it is also necessary to assess the work of each team member [4]. Strategies to address individual assessment build on the students' learning shreds of evidence, be it based on objective observation or subjective perception. The basic types of TWC assessment techniques include the following [5]: 1) simulating events in complex scenarios [6], [7], which are generally used in courses with low number of students; 2) measuring the individual development of TWC based on different scales upon observation of students' work routines and behaviors, [8], [9]; 3) assessing performance of peers and self-assessment [10]-[12]; 4) analyzing objective data obtained from partial results and students' interactions in digital spaces [13], [14]. All these techniques have advantages and disadvantages; for example, the former two are based on observation and limited by the number of students; the third one introduces an important factor of subjectivity and the last one demands a great amount of time and effort from teachers. In addition, these methodologies often require face-to-face activities, and thus may prove too complex in scenarios such as the remote emergency teaching caused by the COVID19 outbreak [15], [16].

This study focuses on the fourth type of techniques (based on analysis of objective data) but aims to overcome the limitations associated with this kind of approach. To do so, we apply a learning analytics tool that facilitates assessment of TWC based on data trails from students' interactions with their teammates in online spaces, irrespective of face-to-face or distance learning.

Prior research on this topic analyzes student interactions in LMS message boards [13], [14]; this information, combined with assessment rubrics applied to partial and final results, helps assess the individual acquisition of TWC [4]. One relevant finding from those research studies is that students are not comfortable with using message boards for interaction with their peers because they consider

* Corresponding author.
E-mail address: mcong@unileon.es

that asynchronous communication is far from the real way in which they usually interact in non-educational contexts. A potential solution to this problem is to move the interaction space to Instant Messaging (IM) tools and applications, which then leads to the development of learning analytics solutions tailored to the characteristics and data structures of these applications.

IM applications have been widely adopted by the general population, but are particularly popular among younger generations. IM enables communication across multiple devices and facilitates synchronous interaction between peers, allowing for different types of messages and data formats (e.g., text, photos, video, voice, etc.) [17]. Younger users favor IM applications over other communication channels, such as phone calls [18], [19]. Additionally, most IM applications do not entail monetary costs and are available for download and use in almost every kind of mobile and non-mobile devices. Currently, the most popular and widely adopted IM application is WhatsApp [20].

Recent studies explored the use of learning analytics tools adapted to the characteristics of WhatsApp [21], [22], but their implementation requires additional data parsing to ensure anonymity, and therefore the collaboration of students and teachers because the phone number must be shared among group members. Other IM tools, such as Telegram, do not have this requirement, and thus may be more suitable to foster collaboration. Telegram offers a free open-source platform without ads, a clean interface and some extra security layers [23]. Further, it incorporates a bot system that facilitates collection and processing of messages without linking them to mobile phone numbers, using internal IDs instead.

This study analyzes the results from the combined use of Telegram and a learning analytics tool for data extraction and processing. More specifically, the research examines collaborative learning settings in two different courses of a Computer Science Degree at the University of León. This examination involves observing student engagement and interactions in Telegram groups and analyzing their relationship with the individual acquisition of TWC. We also compare the results with those of previous cohorts. This study extends [24], which described the learning analytics tool used to collect data in this research and was presented at the TEEM conference 2020. A qualitative assessment of the method based on the feedback received from students complements the quantitative analysis.

The structure of this document is as follows: Section II describes the materials and methods used on the study; Section III presents the results of the analysis; Section IV discusses the main findings and compares them with those from previous research; finally, Section V draws the main conclusions of the study.

## II. Materials and Methods

### A. Materials

Validating a learning analytics tool entails its application in an educational context. In this study, we analyze data from two different Computer Science courses (Operating Systems and Computer Animation) at the bachelor's degree in Computer Science across two different academic years.

- Operating Systems (OS) is a second-year mandatory course delivered to between 100 and 130 students that focuses on the fundamentals of Operating Systems from a practical perspective [21]. Although theoretical concepts are given as lectures, most of the contents are developed as hands-on work. The course assessment consists of the evaluation of theoretical and practical concepts through questionnaires (35 percent of the final grade) and two mandatory assignments to assess the hands-on part (65 percent of the final grade). The latter comprises two assignments: the first one is individual and accounts for 35 percent of the hands-on grade, whereas the second (also called final assignment) is carried out in groups and accounts for the remaining 65 percent of the hands-on grade. Students need to pass both the theoretical and hands-on parts separately to pass the course. This study focuses on the final assignment because of prior success in applying the same methodology (Comprehensive Training Model of the Teamwork Competence, CTMTC; CTMTC is explained in more detail in subsection II.C.2) in project-based learning in the past [25]. The data collected correspond to the 2018-2019 (face-to-face course using Moodle message boards for team communication) and 2020-2021 (blended learning course using Telegram as communication space) academic years.
- Computer Animation (CA) is a third-year elective course where students learn general concepts about design principles and techniques, modeling and three-dimensional animation of objects [24]. The main learning objective of the course is that students experience and learn the concepts involved in all the stages of an audiovisual production project in real contexts. Course contents are divided in three blocks: introduction, animation fundamentals and animation techniques.

  The assessment is based on questionnaires, applied exercises and a final project. Questionnaires (20 percent of the final grade) are used to check students' knowledge and understanding of theoretical concepts. Exercises (20 percent of the final grade) assess students' knowledge about the application of theoretical concepts. The development of an animation project (50 percent of the final grade) is carried out in teams following the CTMTC methodology; the project starts in the first classes and is worked on during the whole semester. The remaining part of the grade corresponds to class attendance. The study examines data collected from two different cohorts: 2018-2019 (face-to face course using Moodle message boards for team communication) and 2019-2020 (shifted to emergency remote teaching due to the COVID19 outbreak and using Telegram as communication tool).

### B. Participants

The sample of the study is described in Table I. Each cell shows the number of students actively participating in the course over the total number of enrolled students. Students in the 2019-2020 Computer Animation and 2020-2021 Operating Systems courses were also given the choice to use message boards instead of Telegram, but all of them chose to use Telegram. Participation in the study was voluntary, and participating students had to explicitly accept and sign a consent form, by which they allowed instructors and the research team to access and analyze their data (the Spanish version of the consent form may be accessed at https://forms.gle/z9dRvSiQZ1PtZkL97). For research purposes, data is anonymized. Students could cancel this agreement at any given moment. Participants were also informed that there were not risks associated with the study, nor any payment due for participation.

From Table I, the number of students in OS doubles the number of students participating in CA, which owes to the mandatory nature of OS and the difference in the year they are taught (OS in second year and CA in third year).

TABLE I. Student Distribution By Course and Cohort

| Course | 2018/19 | 2019/20 | 2020/21 |
|---|---|---|---|
| **OS** | 92/107 | | 105/111 |
| **CA** | 44/52 | 42/56 | |

### C. Applied Methods

This sub-section describes the research methods of the study, the data collection process and a detailed explanation of the methodology followed in the courses (CTMTC), as well as particular aspects of its application in each of the courses.

#### 1. Methods

The research uses a mixed-methods approach [26], combining quantitative and qualitative analysis. The quantitative analysis compares the number of messages and individual TWC acquisition [21] between academic years across courses (which used different tools for team communication purposes) and analyzes the relationship between messages exchanged and individual TWC acquisition. The analysis involves two-sample location tests to find differences between cohorts in both groups, and regression analysis to test for association between messages and individual TWC acquisition.

The qualitative data was gathered from open questionnaires with similar questions about both the methodology (CTMTC) and the digital spaces used for communication. Two questionnaires were delivered: one for courses using message boards (https://forms.gle/6oFeYxEW6HR5Lohv9) and a different one with specific questions for the instant messaging tool (https://forms.gle/51XMURZbEgCAetdd7).

#### 2. Course Methodology

##### a) CTMTC

CTMTC is a methodology designed to develop TWC. CTMTC includes different sequential stages (storming, norming, performing, delivery and documentation), adapted from the project management area as defined by the International Project Management Association (IPMA) [27]. In CTMTC, students develop a project or complex learning activity following sequential phases and working as a team. To complete the different stages, students must make use of different technologies, such as wikis (where they publish their partial results), message boards (where they hold discussions about the project), cloud storage directories (where they upload the deliverables), etc. [13], [28]. The methodology allows for flexibility, as the digital tools may be adapted to different settings [4], [25], [28]-[32]. Using digital tools makes it also possible for instructors to track and analyze students' interactions; for example, instructors may revise the partial results published in the wiki at each phase, or go over what students post to the message boards, the documents they upload to the cloud or publish in a repository, the number of commits in a version control system, etc. These interactions facilitate observation of each team member's participation in every activity and, based on that information and the final work delivered, assessment of the individual acquisition of TWC by each student. However, accessing and analyzing that information is often burdensome, which is why the support of learning analytics tools (such as the one used in this study) become necessary [33], [34].

##### b) CTMTC Application in Operating Systems

CTMTC has been applied to the same assignment (final assignment, total weight of 42.25 percent of the total grade) in both cohorts of the operating systems course. In the 2018-2019 academic year and previous editions of the course, assignation of students to teams was open and free (i.e., students freely choose their teammates), and teams had between three and four members. Each team had to appoint a team coordinator, establish the team's norms and complete the different stages of the CTMTC. Students published the partial results in a Moodle Wiki and interacted with the rest of team members in Moodle message boards. They could also share and publish their results in other virtual repositories, such as Google Drive, Dropbox or GitHub. The rubric described in [32] was used for the assessment of the learning evidence and individual TWC acquisition.

In the 2020-2021 academic year, a different strategy was adopted. While the phases and team assignment did not change, the interaction between team members unfolded in the Telegram IM app. This change had an impact on the rubric, which was no longer applicable because Telegram does not provide information about whether a message is read by a specific student; in addition, the notion of short and long messages is different in IM Tools and message boards. Therefore, a different rubric tailored to the new discussion space [21] was used. Despite this change, both rubrics are similar in that they observe the partial results in the same way.

##### c) CTMTC Application in Computer Animation

The application of CTMTC in the computer animation course focused on the course project, with a weight of 50 percent of the final grade. In the 2018/2019 academic year, the project was developed in teams of between 8 and 9 members, and assignment of students to groups was decided by the instructors. Students published the partial results in Moodle wikis and used message boards to interact and discuss. As in the operating systems course, students made use of cloud-based software (e.g., Google Drive or Dropbox) to share and publish the project's intermediate documents and deliverables. The learning analytics tool analyzed interaction logs from the message boards to help understand and assess students' interactions.

In 2019-2020 academic year, upon realizing that the quality of the final outcomes of the project were subpar, the instructors made some changes, reducing the number of team members to four, and allowing students to freely choose their teammates. Additionally, Telegram was used instead of message boards as interaction space. As in the case of the operating systems course, the original rubric designed for message boards required adaptation to analyze Telegram interactions.

## III. Results

### A. Quantitative Results

As mentioned above, the analysis compares number of messages sent by students and individual acquisition of TWC. Table II summarizes the main descriptive statistics.

TABLE II. Descriptive Statistics

| Course | N | Messages (mean) | Messages (SD) | Indiv. TWC (mean) | Indiv. TWC (SD) |
|---|---|---|---|---|---|
| **OS (18-19)** | 92 | 29.88 | 21.58 | 6.95 | 1.58 |
| **OS (20-21)** | 105 | 167.38 | 169.70 | 7.21 | 2.13 |
| **CA (18-19)** | 44 | 73.45 | 54.56 | 4.98 | 2.80 |
| **CA (19-20)** | 42 | 160.02 | 163.22 | 5.83 | 3.19 |

Fig. 1 shows the interaction plots comparing the results of the analysis of both cohorts. From Fig. 1, there is an overall improvement in individual TWC acquisition and a high increase in the average number of messages posted by students.

To test the significance of the differences, the analysis uses the *ggbetweenstats* function of the *ggstatsplot* package in R [35]. Ggstatsplot combines statistical details and graphical output, making data exploration simpler and faster. Prior to the analysis, we tested for normality of the two variables under study (messages exchanged by each student and individual TWC acquisition) using the Shapiro-Wilk test. Because normality could not be confirmed in the case of

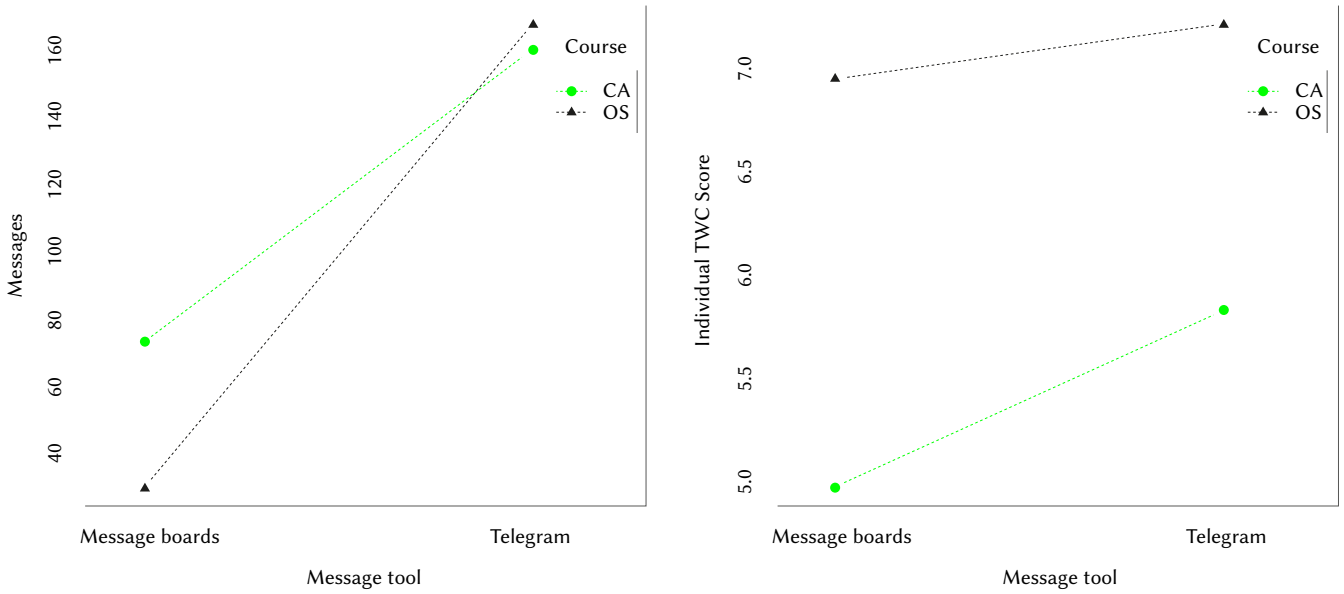



Fig 1. Interaction plots of number of messages (left) and Individual TWC scores (right) across both cohorts in the two courses (CA: Computer Animation; OS: Operating Systems).

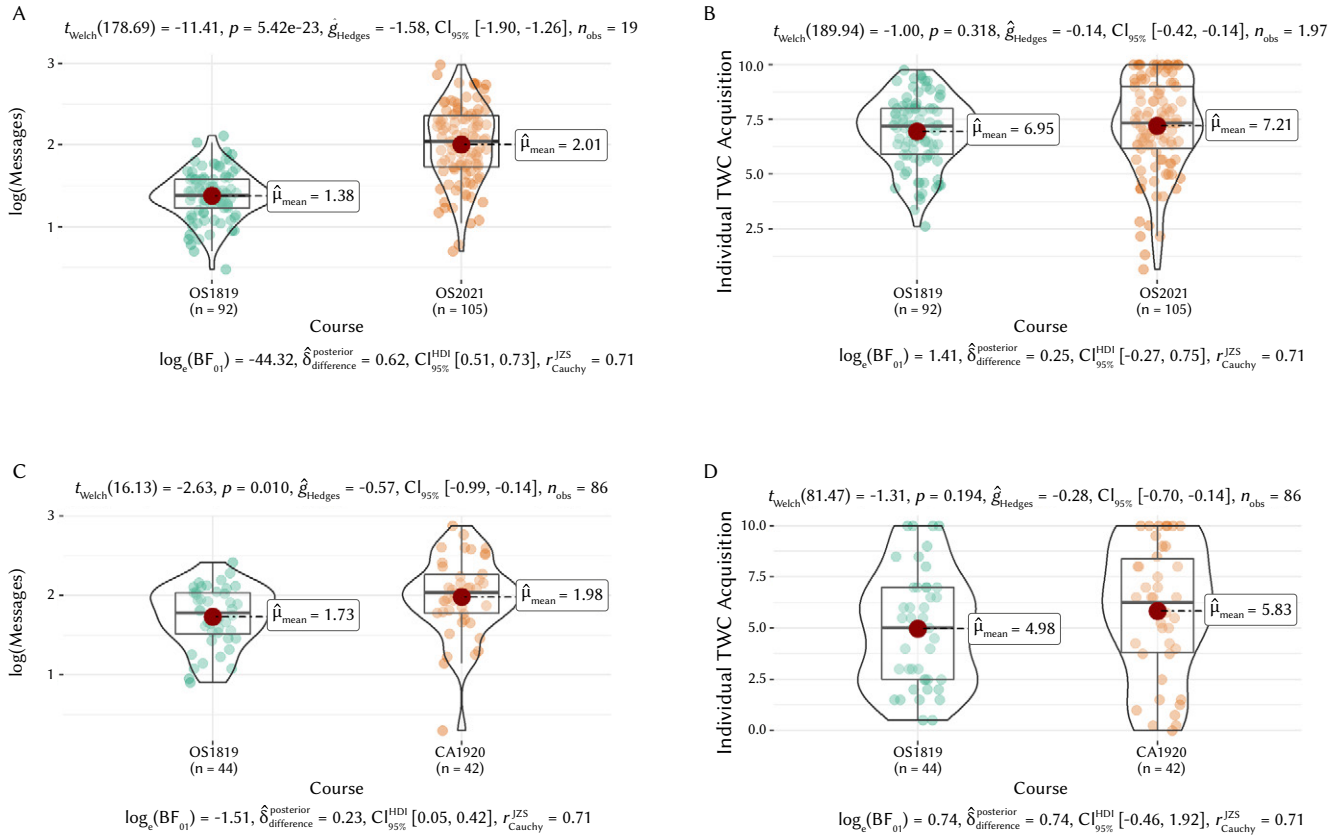



Fig 2. Results of the two samples-location tests of log(Messages) (A, C) and Individual TWC acquisition (B, D) across cohorts of the two courses.

number of messages (p<0.05) , it was necessary to apply a logarithmic transformation of the variable, which then met the normality assumptions. Fig. 2 shows the results of the two-samples location test.

From Fig. 2, the results confirm the significant increase in the number of messages, but the analysis shows no significant increase in individual TWC acquisition.

We then use linear regression to compare the relationship between individual TWC acquisition and messages in each cohort. The analysis includes the interaction effect due to the introduction of Telegram. Table III summarizes the results of the analysis.

From Table III, there is a significant positive relationship between messages and individual TWC acquisition. Further, the results indicate that the influence of number of messages on individual TWC acquisition (i.e., the interaction term) is similar across both cohorts in the two courses.

TABLE III. RESULTS OF THE MULTIPLE LINEAR REGRESSION ANALYSIS

| | Estimate | Std. Error | p-value |
|---|---|---|---|
| **OS (Adj. $R^2$: 0.78)** | | | |
| (Intercept) | 0.65 | 0.44 | 0.14 |
| Messages | 4.57 | 0.31 | 0.00 |
| Telegram | -1.59 | 0.58 | 0.01 |
| Messages*Telegram | -0.51 | 0.36 | 0.16 |
| **CA (Adj. $R^2$: 0.63)** | | | |
| (Intercept) | 5.62 | 1.30 | 0.00 |
| Messages | 6.12 | 0.74 | 0.00 |
| Telegram | 1.46 | 1.74 | 0.40 |
| Messages*Telegram | 1.09 | 0.93 | 0.25 |

### B. Qualitative Results

The qualitative analysis explores students' answers to open questions. Participation was voluntary, and the total number of replies is shown in Table IV. Questions about CTMTC and different software tools other than those relative to interaction and discussion spaces were the same across all courses; questions about Telegram were asked only to students in cohorts that used this IM application. We group the answers by proximity criterion for Q1 (advantages of CTMTC), Q2 (drawbacks of CTMTC), Q3 (additional tools students used to complete the project) and Q4 (advantages of Telegram when compared to asynchronous tools such as message boards). The results are presented in a matrix style, as suggested by [36].

TABLE IV. NUMBER OF REPLIES TO THE QUESTIONNAIRE BY COURSE AND YEAR

| Course | 2018/19 | 2019/20 | 2020/21 |
|---|---|---|---|
| **OS** | 72 | | 90 |
| **CA** | 31 | 22 | |

Table V shows the responses from 40 students (the 10 first answers from each course). In addition, it must be noted that 96.7 percent of the students highlighted the ease of use of the Telegram bot.

## IV. DISCUSSION

The results from this study are in line with previous research. From the quantitative analysis, we can observe similarities across the courses. For instance, the use of Telegram causes an overall significant increase in the number of messages posted by students; the average number of Telegram messages in the computer animation course approximately doubles that of message boards, and in the operating systems course this number is more than five times higher. There are two explanations to this finding:

1. The messages sent through IM apps are generally shorter than those posted to message boards. Therefore, communicating the same content generally entails sending more messages. For reference, a message sent through WhatsApp may be considered long when the number of characters is greater than 40 [37], [38] whereas in message boards long messages contain 150 characters or more [28].
2. Students feel more comfortable with tools they use in their everyday life, provide instant update notifications and are accessible in different devices, particularly mobile devices [39], and therefore tend to use them more often. This explanation is supported by the students' answers to the open question about the use of Telegram.

In both courses and both cohorts of each course, the results show that the relationship between the number of messages and individual TWC acquisition is positive and significant. This result confirms the findings of [4], [40] and shows that engagement and motivation are related to improvement in individual acquisition of TWC [4] and could be also related to an improvement in academic performance [41].

Despite the 3.7 and 17.0 percent increase in individual TWC acquisition in the OS and CA courses, respectively, the analysis cannot confirm whether the use of Telegram leads to significant improvement in TWC acquisition when compared to the use of message boards. Even though further research is necessary to shed light on this finding, we may anticipate some potential causes of these finding, which can be summarized in four explanations:

1. Issues pertaining to the use of mobile devices and instant messaging apps for educational purposes. Although prior research highlights the benefits associated with the use of these tools in learning contexts, especially regarding student interaction [42]-[47], they are also known to cause distraction from the task at hand [48], [49]. It is possible that this effect is also present in the courses under analysis in this study: the results reflect an increase in interactions between team members, but the introduction of Telegram might have led to students not paying the required attention to the group activity. Further analysis investigating the number of multimedia messages, emojis exchanged, as well as discourse analysis using natural language processing could help assess whether interactions were focused on the team activity.
2. Despite the growing use of Telegram, its acceptance is still far from that of other widespread applications, such as WhatsApp [20]. Therefore, students might feel that they are being forced into using Telegram when there is still a practical gap in whether it has already been incorporated to their everyday life. Consequently, its effectiveness may be reduced. Further research is needed to compare the effectiveness of both applications, given their differences in user base and features: user ID, group management, bots, API access, pools, emojis, keyboards or backup processes [50]-[52].
3. Being familiar and proficient with the use of an application for personal use in everyday life does not equate to being able to take advantage of its potential in educational contexts; in other words, being a digital native does not necessarily translate to being a digital learner [53]. As a consequence, proficiency in the use of a tool in a private context may not be associated with an improvement in individual TWC acquisition when using the tool in an educational setting. In addition, even though message exchanges play an important role in individual TWC acquisition, there are other relevant variables influencing individual TWC acquisition, such as other evidence of collaboration activity and leadership [21].
4. Aspects related to COVID19. Prior research suggests that student academic achievement improved under emergency remote teaching during the lockdown period [54], [55]. However, scholarly research has yet to address the effect of the pandemic on team dynamics in project-based learning, which might be potentially hindering TWC development. The computer animation course in the 2019-2020 academic year was given online due to lockdown, and the operating systems course in the 2020-2021 academic year followed a hybrid approach, with half of the groups alternating face-to-face and online sessions every other week. Remote learning might hamper effective teamwork when this skill is yet to be developed by students, and changes in the learning delivery method may reduce both the effectiveness of project-based learning and student motivation, making it more difficult to adequately follow the course. Notably, the change from face-to-face instruction

TABLE V. Selection of Responses to the Open Questions, Classified in Categories
OS1-XX: Operating Systems, 2018-2019 Academic Year; OS2-XX: Operating Systems: 2020-2021 Academic Year
CA1-XX: Computer Animation, 2018-2019 Academic Year; CA2-XX: Computer Animation: 2019-2020 Academic Year

| Course | Q1 (CTMTC Advantages) | Q2 (CTMTC Drawbacks) | Q3 (Other tools) | Telegram |
|---|---|---|---|---|
| **CA1-01** | Working together and planning the tasks | Some people do not know to work in teams | - | - |
| **CA1-02** | Distributing workload | Peers' responsibility | Version control systems | - |
| **CA1-03** | Learning in groups and from others | Group size | WhatsApp, Skype, Telegram | - |
| **CA1-04** | All tasks planned and organized | Initial effort | - | - |
| **CA1-05** | Distributing workload | Use of message boards as communication tool | Other messaging tool | - |
| **CA1-06** | Knowledge Sharing | Agreement with peers | - | - |
| **CA1-07** | Importance of coordination | Effort required by all | Other communication tools | - |
| **CA1-08** | Working in groups | Coordination effort | - | - |
| **CA1-09** | Seeing the whole work | Dependent on how your partners work | Other than message boards | - |
| **CA1-10** | Individual participation in teamwork | Message boards are a bad choice for communication | - | - |
| **CA2-01** | Learning how to work in a team | Coordination effort | - | Useful and easy to use |
| **CA2-02** | Working distribution | Differences in engagement among team members | - | Very straightforward |
| **CA2-03** | Knowing other opinions and solutions | Difficulty to reach consensus | - | Instant messages |
| **CA2-04** | Learning from peers | - | - | Used in my daily life |
| **CA2-05** | Good for planning | Following all the stages | Drive | I would prefer WhatsApp |
| **CA2-06** | Procedure to work in groups | - | Trello | Good for communication |
| **CA2-07** | Distributing responsibilities | Higher workload | Discord | Better than message boards |
| **CA2-08** | Addressing more complex projects as a group | Distribution of responsibility | - | Direct notifications |
| **CA2-09** | Finding better solutions | Effort to work as a group | Skype | Mobile use |
| **CA2-10** | Easier to reach a solution | Involvement and consensus | Discord | Quick to read and answer |
| **OS1-01** | Working as a team | Peers not completing their tasks | - | - |
| **OS1-02** | Addressing big projects | Documenting the progress | A better messaging tool | - |
| **OS1-03** | Learning how to work in a team | Difficulty to coordinate | - | - |
| **OS1-04** | Distributing effort | - | IM tools (e.g., WhatsApp) | - |
| **OS1-05** | Structuring the work and fostering team members participation | Reporting the work done | Telegram | - |
| **OS1-06** | Easy method that facilitates coordination | Mandatory use of message boards | - | - |
| **OS1-07** | Facilitating planning and report of work | Initial understanding of the methodology | WhatsApp | - |
| **OS1-08** | Individual assessment of team members work | Unsuited for large groups | Version control systems | - |
| **OS1-09** | Proper distribution of the workload | Coordination effort | - | - |
| **OS1-10** | Facilitates the coordination between team members | Moodle message boards hindering natural conversation | Instant messaging tools | - |
| **OS2-01** | Easy to apply | - | Discord | Easy to follow with daily life tools |
| **OS2-02** | Workload distribution | - | Discord | Better than message boards |
| **OS2-03** | Reporting the work done as a team | - | Discord, WhatsApp, Notion, Repl.it, GitHub | Comfortable and convenient |
| **OS2-04** | Coordination to work together | - | Skype | Easy to use and mobile |
| **OS2-05** | Structure the work applying a method | Reporting may be hard | Repl.it para, Notion Goodnotes | Better than a message board and multidevice |
| **OS2-06** | Assessing individual contributions | Reporting | Discord | A common and accessible tool |
| **OS2-07** | Working together | Describing the work done | - | Better communication |
| **OS2-08** | Distributing work and assessing it individually | - | Discord | More natural and dynamic communication |
| **OS2-09** | Work organization and tasks distribution | - | - | Instant messages and notifications |
| **OS2-10** | Planning the work | - | - | Easy access to information in mobile phones |

(academic year 2018-2019) to hybrid sessions (academic year 2020-2021) caused a rise in student participation of the course, from an average number of students completing the final assignment of 79 percent in 2016-2017 (not considered in the study) to 88 percent in 2018-2019 and a whopping 97 percent in 2020-2021. In previous editions, students who were not participating or were reluctant to participate generally did not submit the final assignment, but in the 2020-2021 cohort almost all enrolled students submitted the final assignment. Consequently, lower quality projects that were organically filtered out in previous editions were delivered in the most recent course, potentially lowering the average final grade. In addition, other covariates related to decisions about the instructional design might be affecting the results (e.g., different group sizes and instructor-led versus free choice in group configuration) [32].

From the results of the qualitative analysis, the results are in line with findings from previous studies where CTMTC was applied, regardless of the communication tool [29], [32]. All participants mention advantages of the methodology, but one quarter of all students did not find any disadvantage in the application of the methodology. Students highlight benefits associated with project management, planning, workload distribution, reporting and assessing the individual and teamwork. All these aspects are related to teamwork behavior [12], a necessary condition for TWC acquisition.

Regarding disadvantages, 16 percent of students identify the use of message boards as a problem, 40 percent state that more effort, coordination and reporting is necessary when applying CTMTC, and the rest of students point at potential issues related to workload distribution and individual team members not being able to complete the tasks they are assigned in due time.

When observing the use of additional tools, 40 percent of respondents did not seem to need other supporting software or applications. From the remaining 60 percent, three-quarters suggested the adequacy of replacing Moodle message boards (in the cohorts where this tool was used) for IM applications. Alternative digital communication systems used by students include Discord, Skype, and WhatsApp. The remaining responses mention collaborative platforms for document sharing, such as Google Drive; Version Control Systems or Repl.it for code sharing; or Trello and Goodnotes for work organization purposes.

The opinion of the students about Telegram (in the cohorts where it was used) may be summarized in that they find that Telegram is a straightforward tool, simple but powerful enough, with the added benefit that it is compatible with their everyday life and accessible via mobile phones. Students also find positive aspects in bot-based group management and that they do not need to share their personal information.

## V. Conclusion

Teamwork is a highly demanded competence by the labor market and has gained relevance in education. This makes it necessary to assess whether students acquire TWC during their academic education. Assessment of TWC may be performed by observing student interactions when they work in teams. The observation of these interactions may be biased when students communicate through spaces that do not feel natural to them; therefore, for real and natural interaction to occur it is worth considering whether the tools and devices students use are compatible with and integrated in their everyday life. Additionally, assessment of student interactions when working in teams is a time-consuming task, especially in classes with high number of students. This study included the results from the application of a learning analytics tool to facilitate assessment of individual TWC based on student interactions across two courses and two different communication systems (message boards and IM applications).

The main conclusions of the study are that: 1) it is possible to collect and analyze messages of students in IM applications, such as Telegram, as well as to design learning analytics tools with that purpose to facilitate instructors' monitoring and assessment of students; 2) students use and accept IM applications in a more natural way than other systems that have traditionally been in place for communication in learning environments, such as message boards; a benefit of IM applications is that they are multi-device applications that provide students with instant notifications and are more compatible with their lifestyles; 3) CTMTC is flexible enough to be directly applied or easily adapted to different educational contexts and tools; 4) student participation (as per number of messages) has improved with the introduction of Telegram in the courses, which might reflect higher involvement and engagement; 5) the results confirm the strong positive relationship between messages sent and individual TWC acquisition; and 6) the research finds contrarian evidence about the positive influence of IM apps use over message boards as team communication and discussion spaces [21]. The study discusses some of the reasons that could help explain this finding, including the effects of lockdown due to the COVID19 pandemic outbreak, but further research is required to address this issue. Future research should explore other potential educational applications of Telegram in the same courses, beyond the COVID19 context, as well as compare the effects of using different IM applications.

## References


[1] R. Colomo-Palacios, C. Casado-Lumbreras, P. Soto-Acosta, F. J. García-Peñalvo, and E. Tovar-Caro, "Competence gaps in software personnel: A multi-organizational study," *Computers in Human Behavior,* vol. 29, pp. 456-461, March 2013 2013, doi: 10.1016/j.chb.2012.04.021.

[2] D. E. Leidner and S. L. Jarvenpaa, 265-291., "The use of information technology to enhance management school education: A theoretical view.," *MIS quarterly,* vol. 19, no. 3, pp. 265-291, 1995.

[3] D. R. Vogel, R. M. Davison, and R. H. Shroff, "Sociocultural learning: A perspective on GSS-enabled global education," *Communications of the Association for Information Systems,* vol. 7, no. 1, 2001.

[4] Á. Fidalgo-Blanco, M. L. Sein-Echaluce, F. J. García-Peñalvo, and M. Á. Conde, "Using Learning Analytics to improve teamwork assessment," *Computers in Human Behavior,* vol. 47, no. 0, pp. 149-156, 2015, doi: 10.1016/j.chb.2014.11.050.

[5] M. A. Rosen *et al.*, "Tools for evaluating team performance in simulation-based training," *Journal of emergencies, trauma, and shock,* vol. 3, no. 4, pp. 353-359, Oct-Dec 2010, doi: 10.4103/0974-2700.70746.

[6] D. J. Dwyer, R. L. Oser, and E. Salas, "Event-Based Approach to Training (EBAT)," *The International Journal of Aviation Psychology,* vol. 8, no. 3, pp. 209-221, 1998, doi: 10.1207/s15327108ijap0803_3.

[7] N. E. Lane, E. Salas, T. Franz, and R. Oser, "Improving the Measurement of Team Performance: The TARGETs Methodology," *Military Psychology,* vol. 6, no. 1, pp. 47-61, 1994, doi: 10.1207/s15327876mp0601_3.

[8] D. Schwab, H. G. Heneman Iii, and T. A. DeCotiis, *Behaviorally anchored rating scales: A review of the literature.* 2006, pp. 549-562.

[9] A. Frankel, R. Gardner, L. Maynard, A. J. T. J. C. J. o. Q. Kelly, and P. Safety, "Using the communication and teamwork skills (CATS) assessment to measure health care team performance," vol. 33, no. 9, pp. 549-558, 2007.

[10] P. H. Hackbert, "Building Entrepreneurial Teamwork Competencies in Collaborative Learning via Peer Assessments," vol. 1, no. 12, pp. 39-52, 2004.

[11] I. d. l. Ríos-Carmenado, B. Figueroa-Rodríguez, and F. Gómez-Gajardo, "Methodological Proposal for Teamwork Evaluation in the Field of Project Management Training," *Procedia - Social and Behavioral Sciences,* vol. 46, pp. 1664-1672, 2012, doi: https://doi.org/10.1016/j.sbspro.2012.05.358.

[12] K. Tasa, S. Taggar, and G. H. Seijts, "The development of collective efficacy in teams: a multilevel and longitudinal perspective," *Journal of Applied Psychology,* vol. 92, no. 1, pp. 17-27, 2007.

[13] Á. Fidalgo-Blanco, D. Lerís, M. L. Sein-Echaluce, and F. J. García-Peñalvo, "Monitoring Indicators for CTMTC: Comprehensive Training Model

of the Teamwork Competence in Engineering Domain," *International Journal of Engineering Education (IJEE),* vol. 31, no. 3, pp. 829-838, 2015.

[14] D. Lerís, Á. Fidalgo, and M. L. Sein-Echaluce, "A comprehensive training model of the teamwork competence," *International Journal of Learning and Intellectual Capital,* vol. 11, no. 1, pp. 1-19, 2014.

[15] F. J. García-Peñalvo, A. Corell, V. Abella-García, and M. Grande, "Online assessment in higher education in the time of COVID-19," *Education in the Knowledge Society,* vol. 21, Art no. 12, 2020, doi: 10.14201/eks.23013.

[16] F. J. García Peñalvo and A. Corell, "La COVID-19:¿ enzima de la transformación digital de la docencia o reflejo de una crisis metodológica y competencial en la educación superior?," *Campus Virtuales,* vol. 9, no. 2, pp. 83-98, 2020.

[17] C. Lewis and B. Fabos, "Instant messaging, literacies, and social identities," vol. 40, no. 4, pp. 470-501, 2005, doi: doi:10.1598/RRQ.40.4.5.

[18] D. Carnevale, "Email is for old people," *The Chronicle of Higher Education,* vol. 53, no. 7, 2006.

[19] R. Junco and J. Mastrodicasa, *Connecting to the Net.generation: What Higher Education Professionals Need to Know about Today's Students.* US: NASPA, National Association of Student Personnel Administrators, Student Affairs Administrators in Higher Education, 2007.

[20] Statista. "Most popular global mobile messenger apps as of October 2020, based on number of monthly active users." https://www.statista.com/statistics/258749/most-popular-global-mobile-messenger-apps/ (accessed 03/02/2020).

[21] M. Á. Conde, F. J. Rodríguez-Sedano, F. J. Rodríguez-Lera, A. Gutiérrez-Fernández, and Á. M. Guerrero-Higueras, "Assessing the individual acquisition of teamwork competence by exploring students' instant messaging tools use: the WhatsApp case study," *Univ Access Inf Soc,* 2020/11/01 2020, doi: 10.1007/s10209-020-00772-1.

[22] M. Á. Conde, F. J. Rodríguez-Sedano, F. J. Rodríguez-Lera, A. Gutiérrez-Fernández, and Á. M. Guerrero-Higueras, "Analyzing Students' WhatsApp Messages to Evaluate the Individual Acquisition of Teamwork Competence," in *Learning and Collaboration Technologies. Ubiquitous and Virtual Environments for Learning and Collaboration. HCII 2019. Lecture Notes in Computer Science*, Z. P. and I. A. Eds., (Learning and Collaboration Technologies. Ubiquitous and Virtual Environments for Learning and Collaboration, P. Zaphiris and A. Ioannou, Eds. Cham: Springer International Publishing, 2019, pp. 26-36.

[23] T. Sutikno, L. Handayani, D. Stiawan, M. A. Riyadi, and I. M. I. Subroto, "WhatsApp, viber and telegram: Which is the best for instant messaging?," *International Journal of Electrical & Computer Engineering (2088-8708),* vol. 6, no. 3, 2016.

[24] M. Á. Conde, F. J. Rodríguez-Sedano, C. Fernández, A. Gutiérrez-Fernández, L. Fernández-Robles, and M. C. Limas, "A Learning Analytics tool for the analysis of students' Telegram messages in the context of teamwork virtual activities," presented at the Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality, Salamanca, Spain, 2020. [Online]. Available: 10.1145/3434780.3436601.

[25] M. Á. Conde, Á. Hernández-García, F. J. García-Peñalvo, Á. Fidalgo-Blanco, and M. Sein-Echaluce, "Evaluation of the CTMTC Methodology for Assessment of Teamwork Competence Development and Acquisition in Higher Education," in *Learning and Collaboration Technologies: Third International Conference, LCT 2016, Held as Part of HCI International 2016, Toronto, ON, Canada, July 17-22, 2016, Proceedings*, P. Zaphiris and A. Ioannou Eds. Cham: Springer International Publishing, 2016, pp. 201-212.

[26] J. L. Green, G. Camilli, and P. B. Elmore, *Handbook of Complementary Methods in Education Research.* American Educational Research Association by Lawrence Erlbaum Associates, Inc, 2006.

[27] AEIPRO-IPMA. "NCB.- Bases para la competencia en dirección de proyectos."http://www.lpzconsulting.com/images/CP_Trabajo_en_Equipo.pdf (accessed 28/02/2014).

[28] A. Fidalgo, D. Leris, M. L. Sein-Echaluce, and F. J. García-Peñalvo, "Indicadores para el seguimiento de evaluación de la competencia de trabajo en equipo a través del método CTMT," presented at the Congreso Internacional sobre Aprendizaje Innovación y Competitividad - CINAIC 2013, Madrid, Spain, 2013.

[29] M. Á. Conde, F. J. Rodríguez-Sedano, L. Sánchez-González, C. Fernández-Llamas, F. J. Rodríguez-Lera, and V. Matellán-Olivera, "Evaluation of teamwork competence acquisition by using CTMTC methodology and learning analytics techniques," presented at the Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality, Salamanca, Spain, 2016.

[30] M. L. Séin-Echaluce, Á. Fidalgo-Blanco, F. J. García-Peñalvo, and M. Á. Conde, "A Knowledge Management System to Classify Social Educational Resources Within a Subject Using Teamwork Techniques," in *Learning and Collaboration Technologies: Second International Conference, LCT 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Proceedings*, P. Zaphiris and A. Ioannou Eds. Cham: Springer International Publishing, 2015, pp. 510-519.

[31] M. L. Sein-Echaluce, Á. Fidalgo-Blanco, and F. J. García-Peñalvo, "Students' Knowledge Sharing to improve Learning in Engineering Academic Courses.," *International Journal of Engineering Education (IJEE),* vol. 32, no. 2B, pp. 1024-1035, 2016.

[32] M. A. Conde, R. Colomo-Palacios, F. J. García-Peñalvo, and X. Larrucea, "Teamwork assessment in the educational web of data: A learning analytics approach towards ISO 10018," *Telematics and Informatics,* vol. 35, no. 3, pp. 551-563, 2018, doi: 10.1016/j.tele.2017.02.001.

[33] A. Álvarez-Arana, M. Larrañaga-Olagaray, and M. Villamañe-Gironés, "Mejora de los procesos de evaluación mediante analítica visual del aprendizaje," *Education in the Knowledge Society,* vol. 21, Art no. 9, 2020, doi: 10.14201/eks.21554.

[34] A. Martínez-Monés *et al.*, "Achievements and challenges in learning analytics in Spain: The view of SNOLA," *Revista Iberoamericana de Educación a Distancia,* vol. 23, no. 2, pp. 187-212, 2020, doi: 10.5944/ried.23.2.26541.

[35] *Ggstatsplot: "ggplot2" based plots with statistical details.* (2018). Zenodo.

[36] M. B. Miles and A. M. Huberman, *Qualitative Data Analysis: An Expanded Sourcebook.* Sage Publications, 1994.

[37] M. Seufert, A. Schwind, T. Hoßfeld, and P. Tran-Gia, "Analysis of Group-Based Communication in WhatsApp," Cham, 2015: Springer International Publishing, in Mobile Networks and Management, pp. 225-238.

[38] A. Rosenfeld, S. Sina, D. Sarne, O. Avidov, and S. Kraus, "A study of WhatsApp usage patterns and prediction models without message content," *arXiv preprint arXiv:1802.03393,* 2018.

[39] T. Iiyoshi, M. J. Hannafin, and F. Wang, "Cognitive tools and student-centred learning: rethinking tools, functions and applications," *Educational Media International,* vol. 42, no. 4, pp. 281-296, 2005.

[40] Á. F. Agudo-Peregrina, S. Iglesias-Pradas, M. Á. Conde-González, and Á. Hernández-García, "Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning," *Computers in Human Behavior,* vol. 31, no. 0, pp. 542-550, 2// 2014, doi: 10.1016/j.chb.2013.05.031.

[41] C. M. Tavani and S. C. Losh, "Motivation, self-confidence, and expectations as predictors of the academic performances among our high school students," *Child study journal,* vol. 33, no. 3, pp. 141-152, 2003.

[42] A. Edman, F. Andersson, T. Kawnine, and C.-A. Soames, "Informal math coaching by instant messaging: Two case studies of how university students coach K-12 students AU - Hrastinski, Stefan," *Interactive Learning Environments,* vol. 22, no. 1, pp. 84-96, 2014/01/02 2014, doi: 10.1080/10494820.2011.641682.

[43] O. E. Cifuentes and N. H. Lents, "Increasing student-teacher interactions at an urban commuter campus through instant messaging and online office hours," *Electronic Journal of Science Education,* vol. 14, no. 1, 2010.

[44] I. Smit and R. Goede, "WhatsApp with BlackBerry; can messengers be MXit? A philosophical approach to evaluate social networking sites." [Online]. Available: https://repository.nwu.ac.za/handle/10394/13628

[45] S. M. Sweeny, "Writing for the instant messaging and text messaging generation: Using new literacies to support writing instruction," *Journal of Adolescent & Adult Literacy,* vol. 54, no. 2, pp. 121-130, 2010.

[46] S. Lauricella and R. Kay, "Exploring the use of text and instant messaging in higher education classrooms," *Research in Learning Technology,* vol. 21, 2013.

[47] A. Z. Klein, J. C. d.-S.-F.-. Junior, J. V. Vieira-Mattiello-Mattiello-da-Silva, J. L. Victoria-Barbosa, and L. Baldasso, "The Educational Affordances of Mobile Instant Messaging MIM: Results of Whatsapp Used in Higher Education," *International Journal of Distance Education Technologies,* vol. 16, no. 2, pp. 51-64, 2018, doi: 10.4018/ijdet.2018040104.

[48] A. B. Fox, J. Rosen, and M. Crawford, "Distractions, Distractions: Does Instant Messaging Affect College Students' Performance on a Concurrent Reading Comprehension Task?," vol. 12, no. 1, pp. 51-53, 2009, doi: 10.1089/cpb.2008.0107.

[49] R. Junco and S. R. Cotten, "Perceived academic effects of instant messaging use," *Computers & Education,* vol. 56, no. 2, pp. 370-378, 2011/02/01/ 2011, doi: 10.1016/j.compedu.2010.08.020.
[50] Alttop9. "6 Telegram Features that WhatsApp Does Not Have (Updated 2021)." https://cutt.ly/HkfvQaB (accessed 03/02/2021.
[51] MVWConsulting. "Telegram VS Signal, With WhatsApp Comparison Table." https://meganvwalker.com/telegram-vs-signal-with-whatsapp-comparison-table/ (accessed 03/02/2021).
[52] Y. Fernández. "Telegram vs WhatsApp: en qué se parecen y en qué se diferencian ambas aplicaciones." https://www.xataka.com/basics/telegram-vs-whatsapp-en-que-se-parecen-y-en-que-se-diferencian-ambas-aplicaciones (accessed 03/02/2021).
[53] E. E. Gallardo-Echenique, L. Marqués-Molías, M. Bullen, and J.-W. Strijbos, "Let's talk about digital learners in the digital era," *International Review of Research in Open and Distributed Learning,* vol. 16, no. 3, pp. 156-187, 2015, doi: 10.19173/irrodl.v16i3.2196.
[54] T. Gonzalez *et al.*, "Influence of COVID-19 confinement on students' performance in higher education," *PLOS ONE,* vol. 15, no. 10, p. e0239490, 2020, doi: 10.1371/journal.pone.0239490.
[55] S. Iglesias-Pradas, Á. Hernández-García, J. Chaparro-Peláez, and J. L. Prieto, "Emergency remote teaching and students' academic performance in higher education during the COVID-19 pandemic: A case study," *Computers in Human Behavior,* vol. 119, p. 106713, 2021/06/01 2021, doi: 10.1016/j.chb.2021.106713.

Miguel Á. Conde

Miguel Á Conde holds a PhD in Computer Science (2012, University of Salamanca). From 2002 to 2004 he was working in educational environment teaching in several courses related to computers. From 2004 he decided to begin working on software development and eLearning. From 2010 to 2012 he was researching at the University of Salamanca and also working there as a teacher. During 2013 he worked in the Informatics and Communications Service of the University of León and as assistant lecturer in that university. Now he works as an associate professor at the University of León. He is a member of the Robotics research group of the University of León and GRIAL research group of the University of Salamanca. His PhD thesis is focused on the merging of informal, non-formal and formal environments. He has published more than 150 papers about different topics such as eLearning, Service Oriented Architectures, Learning Analytics, Mobile Learning, Human-Computer Interaction, Educational Robotics, etc.

Francisco J. Rodríguez-Sedano

Dr. Francisco J. Rodríguez-Sedano received his Ph.D. degree in intelligent systems for engineering in 2010 from the School of Industrial Engineering and Information Technology at University of León (ULE). Professionally, in 1997 he joined the ULE and currently remains contractually linked to this University as associate professor. In 2014 he joined the Robotics Research Group of the University of León, where he is currently involved in several research projects in the field of social robotics and human-robot interaction. Much of his research interest are accessibility, educational innovation, graphical user interface design and evaluation, human-computer interaction (HCI) and human-robot interaction (HRI).

Ángel Hernández-García

is MSc in Telecommunication Engineering, Master SAP in Integrated Information Systems, and PhD in Information Systems by Universidad Politécnica de Madrid (Spain). He is Associate Professor at the Department of Organization Engineering, Business Administration and Statistics (School of Telecommunication Engineering, Universidad Politécnica de Madrid). He focuses his research on electronic commerce, technology acceptance, social media and learning analytics. He has been guest editor and published research articles in leading international journals.

Alexis Gutiérrez-Fernández

Authors should include their biographies at the end of papers. A typical length for a biography is between 180 and 250 words. The biography can contain the author's educational background, academic and professional life and research expertise. The degrees should be listed indicating institution, country, and year. The photograph is placed at the top left of the biography. The authors can list their research interests. If personal hobbies are included, they will be deleted from the biography.

Ángel M. Guerrero-Higueras

Ángel Manuel Guerrero Higueras has worked as IT engineer at several companies in the private sector from 2000 to 2010 and 2014 to 2016. He also has worked as research assistant in the Atmospheric Physics Group at University of León from 2011 to 2013 and in the Research Institute of Applied Science to Cyber-Security at University of León from 2016 to 2018. He got his Ph.D. at the University of León in 2017. He currently works as Assistant Professor at University of León. His main research interests include robotic software architectures, cyber-security, and learning algorithms applied to robotics.

# Bayesian Knowledge Tracing for Navigation through Marzano's Taxonomy

Francisco Cervantes-Pérez[1]*, Joaquin Navarro-Perales[2], Ana L. Franzoni-Velázquez[3], Luis de la Fuente-Valentín[4]

[1] Universidad Internacional de La Rioja en México, Mexico City (Mexico)
[2] Universidad Nacional Autónoma de México, Mexico City (Mexico)
[3] Instituto Tecnológico Autónomo de México, Mexico City (Mexico)
[4] Universidad Internacional de La Rioja, Logroño (Spain)

Received 27 June 2020 | Accepted 24 April 2021 | Published 14 May 2021

## Abstract

In this paper we propose a theoretical model of an ITS (Intelligent Tutoring Systems) capable of improving and updating computer-aided navigation based on Bloom's taxonomy. For this we use the Bayesian Knowledge Tracing algorithm, performing an adaptive control of the navigation among different levels of cognition in online courses. These levels are defined by a taxonomy of educational objectives with a hierarchical order in terms of the control that some processes have over others, called Marzano's Taxonomy, that takes into account the metacognitive system, responsible for the creation of goals as well as strategies to fulfill them. The main improvements of this proposal are: 1) An adaptive transition between individual assessment questions determined by levels of cognition. 2) A student model based on the initial response of a group of learners which is then adjusted to the ability of each learner. 3) The promotion of metacognitive skills such as goal setting and self-monitoring through the estimation of attempts required to pass the levels. One level of Marzano's taxonomy was left in the hands of the human teacher, clarifying that a differentiation must be made between the tasks in which an ITS can be an important aid and in which it would be more difficult.

## Keywords

Bayesian Knowledge Tracing, Bloom's Taxonomy, Computer-Assisted Instruction, Intelligent Tutoring System, Marzano's Taxonomy.

DOI: 10.9781/ijimai.2021.05.006

## I. Introduction

THE use of computers as helping devices in education started in the early 1960s [1], this was called Computer Assisted Instruction (CAI), which interacted directly with the student, rather than assisting a human professor. A text with questions was shown to the student, who had to provide a brief answer and a set of instructions, and then let the system continue with the next questions. The answers provided by the student were evaluated by the system according to specific patterns. CAIs were frame-oriented systems where, sometimes, students' learning was stimulated while they were engaged in some activity, such as a simulation or a game [2].

During the 70s some Artificial Intelligence (AI) techniques were added to CAI design and were redefined as knowledge-based or Intelligent Computer-Aided Instruction (ICAI) [2]. The teaching strategies were provided by human teachers and written as a set of rules that ICAIs had to apply, to lead students towards an efficient learning process of the subject. In addition, the development of ICAIs allowed the introduction of didactic material to analyze the student's performance after the application of individual tutoring strategies.

Hartley and Sleeman, based on their definition of "intelligent teaching", described that "*a necessary ingredient of an intelligent teaching system is a decision-making algorithm which has specific information about the teaching domain and objectives*" [3]. In addition, they identified two types of components necessary to implement ICAI's decision-making procedures: first, a knowledge representation, for the teaching task and the student model; and second, a control strategy, based on a set of teaching operations and a set of mean-ends guidance rules.

ICAIs were rebranded as ITS (Intelligent Tutoring Systems) and defined as dynamic and adaptive systems for personalized instruction based on students' characteristics and behavior. Their design is the outcome of integrating knowledge from various fields such as: AI, cognitive psychology and educational research. The architecture of an ITS is composed by four modules [4]:

- Domain model: It contains knowledge about the subjects that must be learned. It is also called knowledge model.
- Student model. The structure that stores the student's knowledge status, what the student knows or does not know about the domain.
- Instructional model. It defines the teaching and tutorial strategies. It is also called the teacher model or pedagogical module.
- Interface. It is the media that allows the interaction between the user and the computational system.

Fig. 1 shows the architecture based on these four modules and the way in which the flow of information between them and the user is performed.

* Corresponding author.
E-mail address: francisco.cervantesperez@unir.net

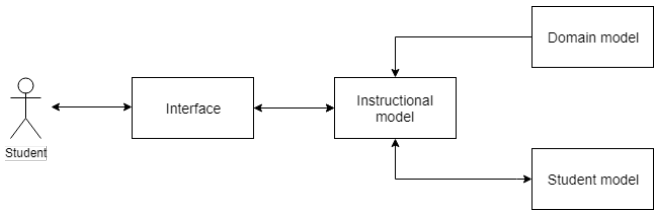



Fig. 1. Architecture of an Intelligent Tutoring System.

From the beginning of the development of ITS there is a very important criticism against them, which consists in affirming that they are not well grounded in a model of learning, and that they seem more motivated by available technology than by educational needs [4]. That is why the authors of this work propose to start from a student advancement system through a taxonomy of educational objectives used in a previous CAI system [5], updating its cognitive foundations and at the same time adding adaptability through a Bayesian model.

The purpose of this work is to propose a theoretical model of an ITS capable of improving and updating computer-aided navigation based on cognitive levels. Our main contribution is the articulation of Marzano's taxonomy of educational objectives, which takes into account the metacognitive system, with the Bayesian Knowledge Tracing algorithm to probabilistically model learners' knowledge.

## II. Previous Work

We classified Intelligent Tutoring Systems into three big groups:

1. Knowledge tracing systems: The systems in the first category model the mastery level of learners and make predictions about it. Some examples are Bayesian Networks to implement a control shared between the students and the machine to track the process of studying linear equations [6], the use of Artificial Neural Networks in children games to determine the right amount of difficulty for each user [7] and Formal Concept Analysis to determine the type of feedback corresponding to each student when solving a given task [8].
2. Conversational agents: Systems in this category use natural language processing to interact with students simulating a human conversation, this is possible because students type text strings either in chat like interfaces or Learning Management Systems sections, and then they are computationally processed. Some examples of the techniques used in these systems and their objectives are semantic web technologies to let students inspect, discuss, and alter their learner models [9], ontologies to model cultural awareness of users through DBpedia database [10], and semantic processing based on conceptual representations to autonomously respond to students' introductions, posted weekly announcements, and answer frequently asked questions [11].
3. Affective tutoring systems (ATS): They are ITS that track the emotional state of student [12]. It is worth mentioning that most of the time a generalized emotional response is estimated, not towards specific problems. ATS are divided into two categories, sensor-based, and sensor-free:

   Sensor based ATS: They use devices such as physiological sensors, pressure sensors, cameras, and eye-trackers. Some examples of these prototypes use photoplethysmographic signals to track reading difficulty [13], a mouse with pressure sensors to measure students' stress [14], facial recognition and the measurement of skin conductance to determine the affective response to concrete problems [15], and eye-tracking to hypermedia environment adaptation [16].

   Sensor-free ATS: They aim to find a correlation between students' emotions and characteristics like interaction logs like number of hints seen, number of hints available, number of skipped tasks, time spent for tasks and time between actions [17] and filled surveys or self-assessment reports, where students report their own feelings, emotions, or mood in a particular learning situation [18]. There are also scopes belonging to this category or to the conversational agents' category and they aim to monitor students' emotions through their interaction with chatbots [19].

Table I shows nine intelligent tutoring systems that are important for the proposal of this work.

As we can see, Bayesian techniques are used to classify learners according to their characteristics and to model their knowledge and performance in an adaptive way. We can also observe that most of the jobs in Table I are based on the level of knowledge of the learners. Our proposal consists of a knowledge tracking system based on a Bayesian model that guides students through specific cognitive levels, to select these levels, we start from the navigation of a CAI system called SAGE.

## III. SAGE

SAGE (Sistema de Apoyo Generalizado para la Enseñanza Individualizada) is a CAI system developed at Tecnológico Autónomo de México (ITAM) [5]. The system has the following characteristics:

### A. Individual Teaching

SAGE allows the learner to select a sequence of topics while meeting the prerequisites for each lesson. This individual teaching approach allows students to take into account variations in their scores and to compare it with the group average, noting their position inside the group.

### B. Content Map

SAGE is based on a content map that organizes subjects from the general to the particular and dependencies are established between the course subjects. Therefore, if the students need to check subjects where they do not need previous knowledge, they will be able to do that, but if they do not have the pre-requirements, the system will not allow them to see the lessons.

### C. Bloom's Taxonomy

Students can progress through lessons solving tests according to the levels of Bloom's taxonomy, this taxonomy operationalizes thinking processes inside a hierarchy which helps to select, describe and evaluate the behaviors that are going to be taught. This is derived from a learning model that considers three domains: cognitive, affective, and psychomotor [29]. The authors proposed six levels for the cognitive level:

- Knowledge: Involves all those behaviors that consist of memorization.
- Comprehension: Understand the message inside the communication process.
- Application: It is the transference of acquired knowledge to similar or almost new situations, this means, to make generalizations.
- Analysis: Split knowledge in their constitutive elements so the relative hierarchy of ideas appears clearly.
- Synthesis: It means the reunion of the elements and parts to form a whole.
- Evaluation: Consists in judging if a determined set of knowledge satisfies or not a specific criterion.

SAGE covers the first four levels of Bloom's taxonomy (knowledge, comprehension, application, and analysis) according to specific types

TABLE I. Examples of intelligent Tutoring Systems

| Authors | Educational field | IA techniques | Purposes of IA techniques | Learner's characteristics |
|---|---|---|---|---|
| Muñoz, Ortiz, Gonzalez, Lopez, and Blobel [20] | Childhood disease management | • Bayesian technique (Bayesian network) | • Define and update student's knowledge level | • Learner's knowledge<br>• Learner's performance |
| Costello [21] | Computer programming | • Data mining technique (Intelligent clustering algorithms)<br>• Condition action rule-based reasoning | • Presenting adaptive learning content<br>• Adaptive recommendation generation<br>• Updating learning styles | • Amalgamated learning style<br>• Learner's preference<br>• Learner's performance |
| Myneni, Narayanan, Rebello, Rouinfar, and Pumtambekar [22] | Physics education | • Bayesian technique (Bayesian network) | • Prediction adaptive learning content<br>• Adaptive feedback and hint generation | • Learner's knowledge<br>• Learner's behavior<br>• Learner's performance |
| Weragama and Reye [23] | Computer programming | • Bayesian-based technique (Bayesian network) | • Determining and updating the student model | • Learner's responses to learning activities |
| Hooshyar, Ahmad, Yousefi, Yusop, and Horng [24] | Computer programming | • Intelligent multi-agent<br>• Bayesian technique (Bayesian network) | • Adaptive feedback and recommendation generation<br>• Levels of knowledge | • Learner's knowledge<br>• Learner's feedback |
| Grawemeyer et al. [25] | Math | • Bayesian technique (Bayesian network classifying and reasoning) | • Classifying the learners affect states<br>• Adaptive feedback generation | • Affect states<br>• Reasoning stage<br>• Learner's interaction |
| El Ghouch, El Mokhtar, and Seghroucheni [26] | Designed for variant courses | • Bayesian technique (Bayesian network classifying) | • Classifying the learners based on learning styles | • Learning style |
| Grivokostopoulou, Perikos, and Hatzilygeroudis [27] | AI curriculum | • Condition action rule-based reasoning (Rule-based expert system)<br>• Data mining technique (decision tree) | • Presenting adaptive exercises<br>• Learners evaluation (prediction of the student performances) | • Learner's knowledge level<br>• Learner's performance |
| Mostafavi and Barnes [28] | Philosophy & Computer science (solving logic proof problems) | • Bayesian-based technique (Bayesian knowledge tracing)<br>• Data mining technique (Cluster-based classification) | • Evaluation and prediction of the learner's performance<br>• Classification of the learners based on their performances | • Student performance<br>• Learner's knowledge |

TABLE II. Correspondence Between Strategies and Cognitive Levels

| Type of question | Knowledge | Comprehension | Application | Analysis | Synthesis | Evaluation |
|---|---|---|---|---|---|---|
| **Brief answer** | ✓ | ✓ | | | | |
| **Completing** | ✓ | ✓ | | | | |
| **Multiple option** | ✓ | ✓ | ✓ | ✓ | | |
| **Matching** | ✓ | ✓ | | | | |
| **Alternative answer** | | | ✓ | ✓ | | |
| **Arranging** | ✓ | | | | | |
| **Essay** | | | ✓ | ✓ | ✓ | ✓ |

of questions, Table II shows the correspondence between evaluation strategies and cognitive levels.

Fig. 2 shows the steps that a learner must carry out in SAGE to select and pass a lesson, and the steps carried out within each of the first four cognitive levels of Bloom's taxonomy (knowledge, comprehension, application, and analysis).

## IV. Proposal

The characteristics and operating principles of the proposed ITS are described below.

### A. Adaptive Learning

The system will allow the navigation path between lessons to automatically adapt to the progress of the learner's skills. For this, the student model starts from the performance of the group to later adapt to individual needs through the Bayesian model.

### B. Bayesian Knowledge Tracing

Transitions between lessons are defined according to the Bayesian Knowledge Tracing algorithm, a tool developed by Anderson and Corbett [30] that modelled the acquisition of knowledge and skills as a Hidden Markov Model, this means, a Markov process with unknown parameters known as hidden states that must be determined from some observable outputs. The unknown parameters are the knowledge and skills that students should possess when their lessons are finished, and the observable outputs are the answers to the evaluation questions, where two options exist: "right" and "wrong".

A personalized sequence of questions is presented to the learner based on probability estimates until the student has mastered each skill. The transition probability represents the odds of a progression between knowledge units, while the emission probability represents the odds of an accurate evaluation. Both probabilities are calculated through a computational procedure that is a variation on one described by Atkinson [31] that employs two learning parameters and

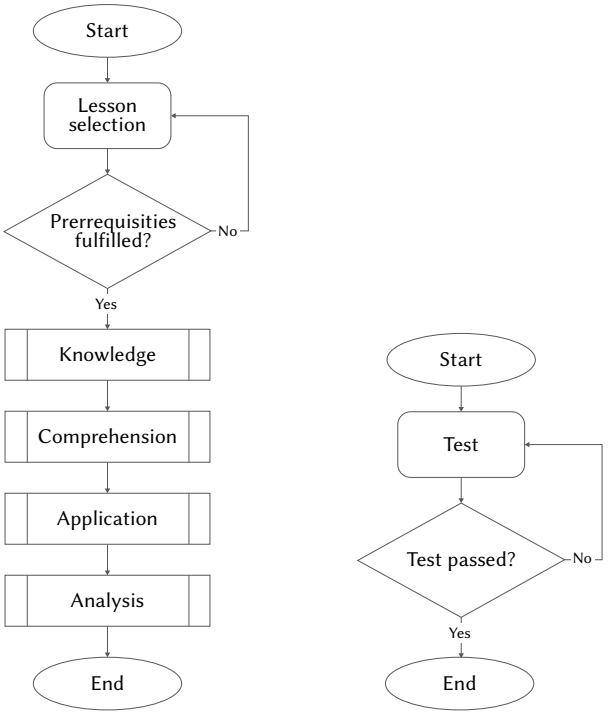



Fig. 2. Navigation in SAGE through the levels of Bloom's taxonomy. a) Workflow of a lesson. b) Subprocess that represents the steps within knowledge, comprehension, application, and analysis.

two performance parameters: Initial Learning or $p(L_0)$ is a learning parameter that indicates the probability that a skill is in the learned state prior to the first opportunity to apply it, Transition or p(T) was described before as the transition probability and it is the second learning parameter. On the other hand, the emission probability is decomposed into two performance parameters: Guess or p(G) is the probability that a student will guess correctly if a skill is in the unlearned state and Slip or p(S) is the probability that a student will make a mistake if a skill is in the learned state. Equations (1), (2), and (3) show the relations between parameters when Initial Learning is updated to $p(L_n)$ [32] where n is the discrete time measure that increases each time an exercise is answered, what is called $Action_n$.

$$P(L_{n-1}|Correct_n) = \frac{P(L_{n-1})*(1-P(S))}{P(L_{n-1})*(1-P(S))+(1-P(L_{n-1}))*(P(G))} \quad (1)$$

$$P(L_{n-1}|Incorrect_n) = \frac{P(L_{n-1})*P(S)}{P(L_{n-1})*P(S)+(1-P(L_{n-1}))*(1-P(G))} \quad (2)$$

$$P(L_n|Action_n) = P(L_{n-1}|Action_n) + \left((1 - P(L_{n-1}|Action_n)) * P(T)\right) \quad (3)$$

Every $Action_n$ has two possible results: correct answer ($Correct_n$) or incorrect answer ($Incorrect_n$). In this way, the parameters $p(L_0)$, p(T), p(G) and p(S) can be calculated from the group's answers, and as each student solves the questions, their $p(L_n)$ will progressively be adjusted depending on whether their individual answers are correct or incorrect.

### C. Marzano's Taxonomy

The questions that allow making transitions between lessons are planned according to Marzano's taxonomy, a taxonomy of educational objectives that proposes a hierarchical order in terms of the control that some processes have over others. The model presents three mental systems: self- system, metacognitive system, and the cognitive system. When the execution of a new task is required, the self-system is responsible for assessing the importance of the task, the probability of success, the present motivation to accomplish it, and the emotional response to the task. Depending on these factors the task is accepted or rejected. When the task is selected, the metacognitive system is responsible for the creation of goals to be achieved, as well as strategies to fulfill these goals. Later, the cognitive system deals with information processing and the analytical operations through four levels of cognition: retrieval, comprehension, analysis, and knowledge utilization [33]. Table III shows correspondence between systems, levels and tasks in Marzano's taxonomy.

The automation of the fourth level of Marzano's taxonomy, which corresponds to knowledge utilization, would require advanced evaluation of texts, therefore the experimental work would be difficult to take into account. The complexity of this level is high for the machine while for the human tutor it is almost intuitive. According to this, the level of utilization of knowledge will be for now in the hands of the human tutors. On the contrary, the Bayesian Knowledge Tracing algorithm will guide the transitions between the retrieval, comprehension, and analysis levels in which it is more feasible to use questionnaires with correct and incorrect answers.

TABLE III. Systems, Levels, and Tasks in Marzano's Taxonomy

| System | Level | Tasks |
|---|---|---|
| **Cognitive** | Retrieval | Retrieval |
| | Comprehension | Integrating, symbolizing |
| | Analysis | Matching, classifying, analyzing errors, generalizing, specifying |
| | Utilization | Decision making, problem solving, experimenting, investigating |
| **Metacognitive** | Metacognitive | Specifying goals, process monitoring, monitoring clarity and accuracy |
| **Self-system** | Self-system | Examining importance, efficacy, emotional response and overall motivation |

The function of the metacognitive system within the algorithm is not a continuation of the cognitive levels, so its role within the knowledge tracing system will be implemented as a function in which the student will be asked how many attempts they will need before the algorithm allows them to go to the next level, thus promoting goal setting and self-monitoring. The system will display the number of attempts that the learner estimated and will advise whether the prediction was correct or not. Fig. 3 shows the steps we propose for a student to pass a lesson, the steps carried out within each of the first three cognitive levels of Marzano's taxonomy (retrieval, comprehension, and analysis), and the step when the learner is asked to estimate the number of attempts it will take to pass.

Regarding the self-system, it is worth mentioning that the adaptive control will let the fastest learners to move forward easily and the slowest learners will be able to move according to their own pace, according to the adjustment of its parameters, avoiding the states of boredom and anxiety that appear when the level of challenge of the activities does not correspond to the student's abilities [34]. In the future, an Affective Tutoring System could be linked to be in charge of monitoring the aspects that correspond to the self-system. We would prefer a sensor-free system to avoid the system being invasive.

## V. Conclusion

The main improvements of our proposal compared to computer-assisted navigation based on cognitive levels are: 1) The adaptive transition between individual questions determined by levels of

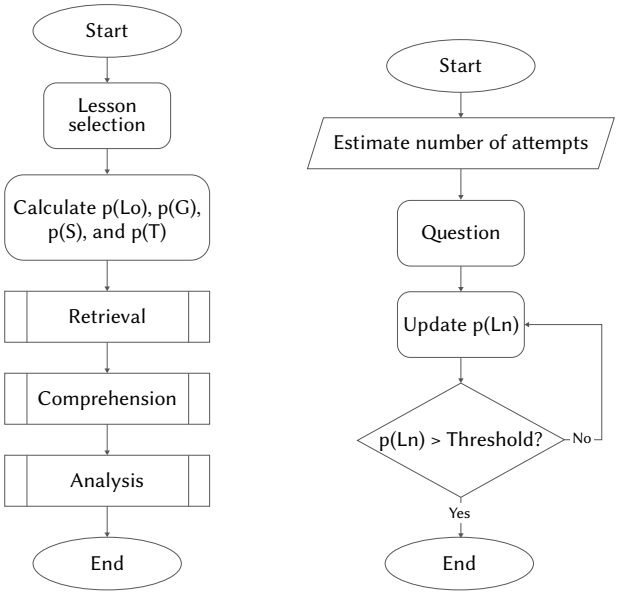



Fig. 3. Proposed navigation through Marzano's taxonomy based on probabilistic parameters. a) Workflow of a lesson. b) Subprocess that represents the steps within retrieval, comprehension, and analysis.

cognition. 2) The possibility of starting the student model based on the general response of the group and adjusting it according to the ability of each learner. 3) The promotion of metacognitive skills such as goal-setting and self-monitoring.

It is worth mentioning that SAGE is based on individualized teaching that at the end of the lessons allows a comparison with the general performance of the group, so its point of comparison is not personalized and could have very different effects on students with different levels of performance. On the contrary, our proposal starts from common parameters that are adjusted in a personalized way, so that the point of comparison is the learners themselves and in this way the level of challenge can be according to their skill level.

In the field of Technology Enhanced Learning, it is important to bear in mind that there are mechanical and repetitive activities that are simple to perform but involve a large amount of time, and complex activities that require considerable effort to perform in a personalized way, especially in large groups. The first can be automated by means of simple resources, as in this case self-grading questionnaires are used for the first three cognitive levels. The second can be assisted by means of artificial intelligence tools, such as personalized transitions between levels according to the Bayesian model. However, there are many activities in which the human teacher has a great advantage over machines and automating them would lead to imprecise and incomplete processes, such as the evaluation of the knowledge utilization level of Marzano's taxonomy. As Sánchez-Prieto *et al.* [35] said, "*it is the moment to reflect on the students' perceptions of being assessed by a non-conscious software entity like a machine learning model or any other artificial intelligence application*".

Clearly delimiting the role of the intelligent tutor system and the human teacher based on a learning model, as in this case, will make it clear that the human teacher is not substitutable and that these types of systems are auxiliary tools for learning. That is, tools can be built to extend teachers' capabilities; for example, in Villagrá-Arnedo *et al.* [36], based on a probabilistic performance prediction system, teachers are given insights on students' learning trends to identify best moments for their intervention.

## Acknowledgement

Work partially funded by the PLeNTaS project, "Proyectos I+D+i 2019", PID2019-111430RB-I00.

## References

[1] H. S. Nwana, "Intelligent tutoring systems: an overview," *Artificial Intelligence Review*, vol. 4, no. 4, pp. 251–277, 1990.

[2] Barr, A. and Feigenbaum, E, *The Handbook of Artificial Intelligence*. Volume 2. HeurisTech Press and William Kaufmann, Inc. Los Altos California, 1982.

[3] Hartley, J. R. and Sleeman, D. H. "Towards more intelligent teaching systems". International Journal of Man-Machine Studies, vol. 5, No.2, pp. 215–236, 1973.

[4] M. Badaracco and L. Martínez, "An intelligent tutoring system architecture for competency-based learning", in International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, 2011, pp. 124-133.

[5] M. Beutelspacher, A. L. Franzoni, and A. Morales, "Sistema de apoyo generalizado para la enseñanza individualizada (SAGE)", Instituto Tecnológico Autónomo de México, México, 1995.

[6] Y. Long and V. Aleven, "Mastery-Oriented Shared Student/System Control Over Problem Selection in a Linear Equation Tutor", in Intelligent Tutoring Systems, vol. 9684, A. Micarelli, J. Stamper, y K. Panourgia, Eds. Cham: Springer International Publishing, 2016, pp. 90-100.

[7] G. Fenza, F. Orciuoli, and D. G. Sampson, "Building Adaptive Tutoring Model Using Artificial Neural Networks and Reinforcement Learning", in 2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT), 2017, pp. 460-462.

[8] G. Fenza and F. Orciuoli, "Building Pedagogical Models by Formal Concept Analysis", in Intelligent Tutoring Systems, vol. 9684, A. Micarelli, J. Stamper, y K. Panourgia, Eds. Cham: Springer International Publishing, 2016, pp. 144-153.

[9] V. Dimitrova and P. Brna, "From Interactive Open Learner Modelling to Intelligent Mentoring: STyLE-OLM and Beyond", International Journal of Artificial Intelligence in Education, vol. 26, no. 1, pp. 332-349, mar. 2016.

[10] R. Denaux, V. Dimitrova, L. Lau, P. Brna, D. Thakker, and C. Steiner, "Employing linked data and dialogue for modelling cultural awareness of a user", in Proceedings of the 19th international conference on Intelligent User Interfaces - IUI '14, Haifa, Israel, 2014, pp. 241-246.

[11] A. K. Goel and L. Polepeddi, "Jill Watson: A Virtual Teaching Assistant for Online Education", p. 21.

[12] M. Magdin, D. Držík, J. Reichel, and S. Koprda, "The Possibilities of Classification of Emotional States Based on User Behavioral Characteristics". International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 4, pp. 97-104. 2020. http://doi.org/10.9781/ijimai.2020.11.010

[13] P. Pham and J. Wang, "Adaptive review for mobile MOOC learning via implicit physiological signal sensing", in Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016, Tokyo, Japan, 2016, pp. 37-44.

[14] J. Hernandez, P. Paredes, A. Roseway, and M. Czerwinski, "Under pressure: sensing stress of computer users", in Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14, Toronto, Ontario, Canada, 2014, pp. 51-60.

[15] A. K. Vail, J. F. Grafsgaard, K. E. Boyer, E. N. Wiebe, and J. C. Lester, "Predicting Learning from Student Affective Response to Tutor Questions", in Intelligent Tutoring Systems, vol. 9684, A. Micarelli, J. Stamper, y K. Panourgia, Eds. Cham: Springer International Publishing, 2016, pp. 154-164.

[16] M. Taub and R. Azevedo, "Using Eye-Tracking to Determine the Impact of Prior Knowledge on Self-Regulated Learning with an Adaptive Hypermedia-Learning Environment", in Intelligent Tutoring Systems, vol. 9684, A. Micarelli, J. Stamper, y K. Panourgia, Eds. Cham: Springer International Publishing, 2016, pp. 34-47.

[17] R. Janning, C. Schatten, and L. Schmidt-Thieme, "Perceived task-difficulty recognition from log-file information for the use in adaptive intelligent tutoring systems", International Journal of Artificial Intelligence in Education, vol. 26, no. 3, pp. 855-876, 2016.

[18] M. Wixon, I. Arroyo, K. Muldner, W. Burleson, D. Rai, y B. Woolf, "The opportunities and limitations of scaling up sensor-free affect detection", in Educational Data Mining 2014, 2014.

[19] M. A. Azim and M. H. Bhuiyan, "Text to Emotion Extraction Using Supervised Machine Learning Techniques", Telkomnika, vol. 16, no. 3, 2018.

[20] D. C. Muñoz, A. Ortiz, C. Gonzalez, D. M. Lopez, and B. Blobel. "Effective e-learning for health professional and medical students: The experience with SIAS-intelligent tutoring system". Studies in Health Technology and Informatics, Vol. 156, pp. 89–102. 2010.

[21] R. Costello. "Adaptive intelligent personalised learning (AIPL) environment (U621351 Ph.D.)", University of Hull (United Kingdom), Ann Arbor. ProQuest Dissertations and Theses A&I; ProQuest Dissertations & Theses Global database. 2012.

[22] L. S. Myneni, N. H. Narayanan, S. Rebello, A. Rouinfar, and S. Pumtambekar. "An interactive and intelligent learning system for physics education". IEEE Transactions on Learning Technologies, Vol. 6, no. 3, pp. 228–239. doi:10.1109/TLT.2013.26

[23] D. Weragama and J. Reye. "Analysing student programs in the PHP intelligent tutoring system". International Journal of Artificial Intelligence in Education, vol. 24, no. 2, pp. 162–188. 2014.

[24] D. Hooshyar, R. B. Ahmad, M. Yousefi, F. D. Yusop, and S.J. Horng. "A flowchart-based intelligent tutoring system for improving problem-solving skills of novice programmers". Journal of Computer Assisted Learning, vol. 31(4), pp. 345–361. 2015.

[25] B. Grawemeyer, M. Mavrikis, W. Holmes, G. S. Sergio, M. Wiedmann and N. Rummel. "Affecting Off-task Behaviour: How affect-aware feedback can improve student learning". ACM international Conference Proceeding Series. 2016.

[26] N. El Ghouch, E.-N. El Mokhtar and Y. Z. Seghroucheni. "Analysing the outcome of a learning process conducted within the system ALS_CORR [LP]". International Journal of Emerging Technologies in Learning, vol. 12, no. 3, pp. 43–56. 2017.

[27] F. Grivokostopoulou, I. Perikos and I. Hatzilygeroudis. "An educational system for learning search algorithms and Automatically Assessing student performance". International Journal of Artificial Intelligence in Education, vol. 27, no. 1, pp. 207– 240. doi:10.1007/s40593-016-0116-x. 2017.

[28] B. Mostafavi, and T. Barnes. "Evolution of an intelligent deductive logic tutor using data-driven elements". International Journal of Artificial Intelligence in Education, vol. 27, no. 1, pp. 5–36. 2017.

[29] B. S. Bloom, "Taxonomy of educational objectives. Vol. 1: Cognitive domain", N. Y. McKay, pp. 20-24, 1956.

[30] A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge", User Modeling and User-Adapted Interaction, vol. 4, no. 4, pp. 253-278, 1995.

[31] R. C. Atkinson and J. A. Paulson, "An approach to the psychology of instruction", Psychol. Bull., vol. 78, no. 1, pp. 49-61, 1972.

[32] R. S. J. d. Baker, A. T. Corbett, and V. Aleven, "More Accurate Student Modeling through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing", in Intelligent Tutoring Systems, vol. 5091, B. P. Woolf, E. Aïmeur, R. Nkambou, y S. Lajoie, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 406-415.

[33] R. J. Marzano and J. S. Kendall, The new taxonomy of educational objectives. Corwin Press, 2006.

[34] M. Csikszentmihalyi, *Applications of Flow in Human Development and Education*. Dordrecht: Springer Netherlands, 2014.

[35] J.C. Sánchez-Prieto, J Cruz-Benito, R. Therón, and F. García-Peñalvo, "Assessed by Machines: Development of a TAM-Based Tool to Measure AI-based Assessment Acceptance Among Students". International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 4, pp. 80-86, 2020. http://doi.org/10.9781/ijimai.2020.11.009

[36] C.J. Villagrá-Arnedo, F.J. Gallego-Durán, F. Llorens-Largo, R. Satorre-Cuerda, P. Compañ-Rosique, and R. Molina-Carmona, "Time-Dependent Performance Prediction System for Early Insight in Learning Trends". International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 2, pp. 112-124, 2020. http://doi.org/10.9781/ijimai.2020.05.006

Francisco Cervantes-Pérez

Francisco Cervantes-Perez. He received his B.E. in Mechanical Electrical Engineering and a M.E. in Electrical Engineering, both from the National Autonomous University of Mexico (UNAM), and his Ph.D. in Computer and Information Sciences from the University of Massachusetts at Amherst, Mass, USA. He is a member of the National System of Researchers in Mexico, the Mexican Academy for Informatics, the Mexican Academy for Computer Science, the Mexican Society for Artificial Intelligence, among other. His main research interests are in Computational Neuroscience, Artificial Intelligence, Ecological Robotics, and Technologies for On-line Education. Currently, he is the Rector of the International University of La Rioja in Mexico.

Joaquin Navarro-Perales

Joaquin Navarro-Perales is an academic technician in Coordinación de Universidad Abierta, Innovación Educativa y Educación a Distancia (CUAIEED), at Universidad Nacional Autónoma de México (UNAM). He has a master's degree in Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional (CINVESTAV) and received his bachelor's degree in Biomedical Engineering from Universidad de Guadalajara (UDG). Currently, he is a PhD student in Computer Science at Universidad Internacional de la Rioja (UNIR), and he is working towards a bachelor's degree in Pedagogy at Universidad Nacional Autónoma de México (UNAM). His research interests include Intelligent Tutoring Systems, Self-Regulated Learning, and Pedagogical Conversational Agents.

Ana L. Franzoni-Velázquez

Dr. Franzoni is a full-time professor and director of the Computer Engineering degree program at the Instituto Tecnológico Autónomo de México (ITAM). She has a PhD in Knowledge Engineering and Information Systems from the Université de Technologie de Troyes (UTT) and TELECOM & Management SudParis (France). Her Master's degree is in Information Technology and Management from ITAM and she also has a master's degree in Networks and Information Systems for Companies from the École Nationale Supérieure des Télécommunications de Bretagne (ENSTB) (France). Her undergraduate degree is in Computer Engineering, at ITAM. Dr. Franzoni specializes in Technology in Education, learning environments, computer skills, intelligent tutorial systems and learning analytics. She has produced several academic publications with international arbitration that have appeared in journals, book chapters and conferences. She also is an editorial board member in different journals. Dr. Franzoni has been a visiting professor at San Jose State University (California) and at TELECOM & Management SudParis (France). Dr. Franzoni is the director of the laboratories for mobile devices, Web and videogames at ITAM and in this context she develops links with companies involved in mobile application design, video games and emerging technologies. Dr. Franzoni is a member of the National Association of Educational Institutions in Information Technology, A.C. (ANIEI), evaluator of the National Council of Accreditation in Computer Science A.C. (CONAIC), member of the Mexican Society of Computing in Education, A.C.(SOMECE), member 259 of the Mexican Academy of Information Technology, A.C.(AMIAC), member of the Mexican Association of the Information Technology Industry (AMITI).

Luis de la Fuente Valentín

Dr. Luis de-la-Fuente-Valentín is a full-time associate professor at Universidad Internacional de La Rioja (UNIR), at the School of Engineering and Technology. Before joining this institution, he obtained his degree in Telecommunication Engineering in 2005 and then he started a research grant at Universidad Carlos III de Madrid, where he obtained his PhD in 2011. He leads the Data Science research group, with research topics focused on artificial intelligence, machine learning techniques, natural language processing and data centered applications. He has authored more than 40 papers and participated in several Spanish and European public funded projects, one of them as investigator in charge. His research experience focuses on Technology Enhanced Learning, Learning Analytics and Natural Language Processing applied to the educational field.

LA UNIVERSIDAD
EN INTERNET

Rectorado
Avenida de la Paz, 137
26006 Logroño (La Rioja)
t (+34) 941 21 02 11

www.unir.net