International Journal of Interactive Multimedia and Artificial Intelligence

December 2022, Vol. VII, Number 7
ISSN: 1989-1660

unir LA UNIVERSIDAD EN INTERNET

"The entire effort of artificial intelligence is essentially a fight against computers' rigidity."
Douglas Hofstadter

INTERNATIONAL JOURNAL OF INTERACTIVE MULTIMEDIA AND ARTIFICIAL INTELLIGENCE

ISSN: 1989-1660 –VOL. 7, NUMBER 7

EDITORIAL TEAM

Editor's Note

The International Journal of Interactive Multimedia and Artificial Intelligence – IJIMAI –provides an interdisciplinary forum in which scientists and professionals can share their research results and report new advances in Artificial Intelligence (AI) tools or tools that use AI with interactive multimedia techniques. The present regular issue includes 13 articles. The first block of articles deals with problems related to images as diverse as the artificial generation of images or the optimization of their storage and transmission through compression techniques. The applications are very diverse, including the identification of forgeries, tumors or even misplaced face masks. Another block contains only one paper on speech recognition targeted on specific users suffering from dysarthria. Other block of two articles focuses on the education field problems of automation of teachers' certification processes or prediction of students' academic failure. Last block of articles covers services and products, commerce, marketing and user experience issues, as well as the ethical implications of AI.

In the last years, outstanding results have been obtained by synthetic image generation algorithms applying deep learning based approaches. These can create original works of art or others astonishingly faithful to the artists' works. The first article by Fraile-Narváez et al. questions the possibility of using AI to detect forgeries created by AI algorithms, in order to protect works and their attribution. The authors propose a convolutional neural network (CNN) that is trained with a dataset of paintings by Rembrandt and other 17th century Dutch painters, with similar artistic styles. This is key to detect forgeries when paintings are remarkable similar. Their experiments showed that it was possible to create a deep learning algorithm capable of detecting false images generated by AI algorithms, specifically by Dall-e 2, with a high degree of accuracy.

The following article also proposes a CNN to solve a today's problem very different from the previous one. Since the coronavirus pandemic, face masks have been a common defense method worldwide to protect from respiratory diseases. Bhaik et al. propose to use the light-weighted neural network MobileNetV2 to detect people who are not wearing these masks properly. Their network is based on the concept of transfer learning and is trained on a self-made dataset of images, achieving an accuracy higher than 90%.

Next, Arif et al. propose an adaptive deep learning model for the detection of multi-fog types of images. Detection of objects in foggy weather condition is important for many applications such as autonomous driving or video surveillance. Foggy scenes are of different types based on fog density level and fog type, and detecting these types as a pre-processing step can enhance the detection results. Their experiment shows a 96% classification accuracy rate. Besides, the authors provide a dataset of multi-fog scenes due to the lack of publicly available datasets of inhomogeneous, homogenous, dark, and sky foggy scenes.

Unfortunately, another today's relevant problem is tumor identification. During the last years, many works have been undertaken on image segmentation for medical problems, as identification of brain tumors, using Magnetic Resonance (MR) images. Khemchandani et al. propose a classifier, based on CNN, to categorize the brain tumor from the MRI image segments. The CNN is optimized by using an algorithm that combines particle swarm optimization (PSO) and the imperialist colony algorithm. Also, they propose the scatter local neighborhood structure description to capture textural characteristics to support accurate tumor categorization. The proposed method obtained a maximum accuracy of 0.965 during the experimentation utilizing the Multimodal Brain Tumor Image Segmentation Benchmark (BRATS) database.

As previous works show, the applications of images are many, and there is an increased demand for their storage and transmission. Image compression algorithms contribute to their efficient storage and transmission. In their article, Kaur et al. propose a new technique for optimizing image compression using Fast Fourier Transform (FFT) and Intelligent Water Drop (IWD) algorithm. The aim is to increase compression while keeping its best possible quality. The IWD is used to optimize de FFT threshold values. The Structural Similarity Index Measure (SSIM) is used to estimate the perceived image quality, obtaining further understanding of the compression problem and promising results.

Some diseases, as the previous mentioned brain tumor, cause dysarthria, which reduces the speech quality of a person by affecting the speech production system. Therefore, the automatic speech recognition systems degrade when the speaker suffers dysarthria. The following paper by Sahu et al., proposes a four-level discrete wavelet transform (DWT) decomposition to capture the sub-band information of the speech signal. Then, using the Inverse IDWT, the signals are reconstructed and the log filterbank energies are computed by analyzing the short-term discrete Fourier transform magnitude spectra of each reconstructed speech signal. For each analysis frame, the log filterbank energies obtained across all reconstructed speech signals are pooled together, and discrete cosine transform is performed to represent the cepstral feature, that is the discrete wavelet transform reconstructed (DWTR)- Mel frequency cepstral coefficient (MFCC) in this study. Given the results in the experiments, the authors propose a two-stage classification approach by using MFCC and DWTR-MFCC features, improving classification accuracy.

The following work is related to web information purchasers, whose consumption patterns are changing, being more focused on psychological satisfaction than price satisfaction. In multimedia content searches, psychological satisfaction can be improved through searching by mood or emotion rather than text content. Social Networks Services (SNS) use this type of searches based on a folksonomy, for example, but there are problems with synonyms. To solve the problem of synonyms in their previous study, Moon et al. represent the mood in multimedia content with arousal and valence (AV) in Thayer's two-dimensional model. Although some problems of synonyms were solved, the retrieval performance of the previous study was less than that of a keyword-based method. In their present study, the authors propose a new method where the mood of multimedia content is represented with a fuzzy set of 12 moods of the Thayer model. The experiments show that the proposed method is superior to other two methods, one based on AV value and the other based on keywords.

During the last years, the use of learning management systems (LMS) as a complement to face-to-face classes has spread a lot. Many universities study the usage of LMS by teachers to certificate and evaluate their competence in technology-based learning. To automate this process, in the next paper, Regueras et al. present an expert system that automatically classifies courses and certifies the teachers' competence in LMS from the data in their logs. Firstly, clustering helps define the classification scheme, which is used to define the rules used to classify courses. The scheme and rules have been obtained from data of 3303 courses and two million interactive events. The system has been tested with real data and the results have been successfully

DOI: 10.9781/ijimai.2022.11.009

validated against human experts. These experts valued very positively to have a tool that can automate the process as they find very difficult to manually classify many of the courses.

Also, in the field of education, the next study by Rincon-Flores et al. aims at improving the learning-process and reducing academic failure in two Physics courses. Specifically, they use K-nearest neighbor and random forest algorithms to predict academic performance. In their experiments they find differences between the first and second term evaluations, obtaining not very accurate predictions in the first term evaluation. However, the accuracy improves in the second term evaluation, as datasets grows. Based on the research results, the algorithm delivered a forecast of the group performance in general. Therefore, the algorithm can be used by the instructor to design and implement adaptative measures during the course.

In the recent years, the integration of work and family life is difficult since workers use technologies to work at home. This is stressful for workers who work anytime and anywhere. Therefore, a balanced work-life is important to improve mental and physical health condition of workers. Majumder et al. propose a monitoring web-based tool called the 'Wheel of Life.' Its interface helps tune various important factors, such as business life, creativity, social life, love, and life purpose, and provides multiple recommendations. Users can choose any of those recommendations and improve the living areas accordingly.

User experience is key for the success of a product or service. Today's users demand for high level of satisfaction, doing tasks efficiently and, moreover, they expect hedonic qualities not directly target-oriented. Agile methods are used frequently to develop products reducing development time. Requirements are typically written in user stories. In the next paper, Hinderks et al. propose a method called UX Poker to estimate the impact of a user story on user experience before the development of the product or service. Their results show that UX Poker can be implemented in real-life applications, providing essential insights for the agile team.

The use of AI is increasing rapidly, covering many different areas as health, marketing or education, as the present issue shows. Although the advantages of AI are recognized, there is a growing concern on the trust in systems. The European Union (EU) states that one of the four basic pillars on the development of AI is the development of ethical and trustworthy AI. In their paper, López Rivero et al. present an empirical study on the perception of the ethical challenges of artificial intelligence groups in the classification made by the EU. Authors seek to identify the ethical principles that cause the greatest concern among the population, believing that the study is a starting point for an informed debate on the ethical implications of AI based on the classification of ethical principles made by the EU.

The mentioned growth of use of AI and digital services have accelerated the opportunities of marketing intelligence. Supported by mobile and personalized marketing services, customer experiences have been optimized. However, on the other hand, there is also some deterioration caused by annoying marketing communication in real-time or programmatic advertisement, for example. Therefore, the question if marketing intelligence means service boom or bust of marketing arises. In the last paper of this issue, Lies elaborates the boom and bust aspects of marketing intelligence through a literature review, concluding that the question whether marketing 4.0 means the boom or bust of marketing remains open.

Dr. Rubén González Crespo
Editor-in-Chief
Universidad Internacional de La Rioja

TABLE OF CONTENTS

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.*

Painting Authorship and Forgery Detection Challenges with AI Image Generation Algorithms: Rembrandt and 17th Century Dutch Painters as a Case Study

Marcelo Fraile-Narváez*, Ismael Sagredo-Olivenza, Nadia McGowan

Universidad Internacional de La Rioja (Spain)

Received 2 October 2022 | Accepted 15 November 2022 | Published 28 November 2022

Abstract

Image authorship attribution presents many challenges and difficulties which have increased with the capabilities presented by synthetic image generation through different artificial intelligence algorithms available today. The hypothesis in this research considers the possibility of using artificial intelligence as a tool to detect forgeries through the usage of a deep learning algorithm. The proposed algorithm was trained using a dataset comprised of paintings by Rembrandt and other 17th century Dutch painters. Three experiments were performed with the proposed algorithm. The first was to build a classifier able to ascertain whether a painting belongs to the Rembrandt or non-Rembrandt category, depending on whether it was painted by this author or not. The second tests included other 17th century painters in four categories. Artworks could be classified as Rembrandt, Eeckhout, Leveck or other Dutch painters. The third experiment used paintings generated by Dall-e 2 and attempted to classify them using the prior categories. Experiments confirmed the hypothesis with best executions reaching accuracy rates of more than 90%. Future research with extended datasets and improved image resolution are suggested to improve the obtained results.

Keywords

Artificial Intelligence, Authentication, Image Generation, Machine Learning, Neural Network.

DOI: 10.9781/ijimai.2022.11.005

I. Introduction

SYNTHETIC image generation algorithms based on deep learning have become increasingly popular in recent years. Advances in this field have brough surprising results, such as new neural networks capable of generating images with incredible precision. Among them are Dall-e 2 [1], based on CLIP [2], Midjouney [3] or Imagen, by Google [4]. These algorithms are focused on the generation of artificial images according to a set of keywords. They combine language understanding, classification, and image generation systems.

Essentially, we are faced with a new technology with significant potential, but which has also led to concern among contemporary artists as they fear their work may be endangered by these algorithms. There has even been speculation regarding creating new works by deceases artists [5]. An example of this is the case presented by AI expert Carlos Santana [6], who published on social media some results obtained using Dall-E 2 and famous artworks, creating painting astonishingly faithful to the artists' works. AI synthetic image generation has generated controversy among artists. Some extremes dismiss this technology, considering the images it generates lack artistic value [7] while others fully embrace them [8].

This paper would like to question the possibility of using AI as a tool to protect artworks and their attribution by using it as a tool to detect possible forgeries created by AI algorithms. The working hypothesis of this research is that it is possible to formulate a deep learning algorithm capable of detecting with a high degree of accuracy whether an image is authentic, or a digital falsification created by an AI. Such an algorithm would have other potential applications, such as, for example, supporting specialists in the first stage of attributing artworks whose creators are anonymous or unknown.

The question that drives this article is not foreign to academic research. Similar research has been carried out regarding deepfakes [9]. These are videos where deep learning models are used to substitute the face of a person (usually well-known) for someone else's. These videos can have humorous intent but can also harbor malicious purposes [10]-[14]. The possible implications of attacks with modified images to hack learning systems that have images as an input, such as adversarial attacks in classification systems, have also been studied [15].

A new reality is emerging, where machines are now able to learn, to compose music, and to paint like Rembrandt. Through a deep learning algorithm and facial recognition, with devices within everyone's reach, it is possible for a program to create its own artwork. This process seems to blur the boundaries between art and technology.

A relevant example of the use of AI art creation technology is that of The Next Rembrandt. This project aimed to generate a painting by the Dutch Baroque painter Rembrandt van Rijn (1606-1669) that was indistinguishable from his originals. This project involved Microsoft, the Dutch bank ING, the advertising agency J. Walker Thompson, the Technical University of Delft, the Mauritius Museum, and the Museum Het Rembrandthuis. In addition to the artificial intelligence used to

* Corresponding author.
E-mail address: marcelo.fraile@unir.net

create the painting, the project also sought to recreate the techniques used by the artist using a 3D printing technology that replicated the texture of oil painting.

Specialists question whether it would be possible to establish the authenticity of a work of art in the midst of conflicting expert opinions [16]-[17].

Two considerations must be taken into account when working with these paintings. Firstly, many artists benefited from the help of their disciples. At times, they left in their hands the execution of some details of their work. One well-known case is that of Leonardo Da Vinci and his intervention in different pieces by the master Verrocchio, to the point where today it is debated whether the latter intervened in some of these paintings. Secondly, many unsigned images have been attributed to painters such as Rembrandt and later studies have dismissed or questioned these attributions.

Secondary studies on the work of art are important to ascertain attribution. An artwork is comprised of several layers, other than the final painting seen. There is a primer, guides or sketches drawn on the canvas and modifications made during the painting process which can encompass variations in the background or other details.

An interesting project in this field is that developed by the Zhejiang University of Technology in China, which seeks to create a machine learning model capable of describing and classifying works of art by genre and style. In their results [18] the authors highlighted the importance of using convolutional neural networks to classify art.

This study experimented with seven different models applied to three different datasets under the same experimental setup. The algorithm initially categorizes images according to style and genre, and then classifies them by looking for similarities. A growing number of projects use CNN models to solve classification problems. However, training CNNs requires datasets containing a large number of labeled entries. Their success depends on the availability of large datasets such as ImageNet. For this project, the authors used Cafenet, a slightly modified version of the AlexNet model, to evaluate the fine-tuning process using five pre-trained networks.

Their dataset images were obtained from WikiArt, an organized collection of more than 80,000 images, with more than 1000 artists, 27 different styles and 45 genres separated into different categories. The size of each sample was set at 450 x 450, neither too small so as not to limit the analysis of fine details, nor too large so as not to overfit the CNN with the training data. This resulted in a success rate of more than 90 %. The CNN initially took only color information to classify the painting, later it included spatial information to help the model distinguish portraits from landscapes. It does, however, present problems to identify individual painters from styles.

Steven Frank, in turn, developed a CNN capable of identifying painters such as Picasso, Van Gogh or Rembrandt. It generates a probably map through the division of an imagen divided into a mosaic of small square fragments that can be handled by the CNN, while increasing the number of images used for its training [19]. In his research, Frank developed a CNN to identify authentic Rembrandts from forgeries. He selected 50 portraits by Rembrandt and 50 by randomly selected artists. Some had a very similar style to his and others, while similar, were clearly distinguishable from his work. This choice was made due to the fact that if they were too much alike, the CCN would over-fit and not generalize its training.

Other machine learning techniques have also been used for the classification of works of art, as for example in the work of Wu [20], Xu [21] or Blessing and Wen [22]. In the latter, the authors try to perform two-by-two classifiers among a set of painters including Cezanne, Dali, Durer, Monet, Picasso, Rembrandt, and Van Gogh. As feature extractors, they used different algorithms such as GIST [23], HOG2x2 [24], Dense SIFT [25], etc. And for classification Support Vector Machine was used, with results ranging between 90% and 95%.

Narang and Soriano [26] used the Gray-Level Co-occurrence Matrix (GLCM) to extract the characteristics of a painting and to classify a neural network or a Support Vector Machine with a gaussian kernel, obtaining results of around 83% and 85% accuracy in the detection of paintings created by Juan Luna. For training, this project used 13 high resolution paintings created by Luna, and another 13 by other Filipino artists.

In this study, a CNN will be trained with a dataset developed based on similar artistic styles. This is intended to make the network more sensitive to small perturbations in the artists' styles. These considerations are key to detect forgeries more effectively between paintings that are already remarkably similar. Thus, seeking to focus this research, artworks by seventeenth-century Dutch artists contemporary with Rembrandt's academy are selected.

Rembrandt was chosen as the main painter of this study due to two conditioning factors: firstly, the number of works sufficiently large to build an acceptable training corpus and, secondly, having enough imitators, disciples, attributed paintings and the existence of The Next Rembrandt, a painting generated using Microsoft's AI. All these resources implied an abundant amount of information to assess the algorithm.

II. Methods

To test our hypothesis, a machine learning model capable of identifying Rembrandt's paintings among authors sharing aesthetic similarities was created to identify the details that differentiate Rembrandt from his imitators, in the hope that the network would also learn to distinguish forgeries. To this end, and drawing on the literature, a convolutional network was chosen as an image feature extractor. Specifically, a feature extractor encompassed within the MobileNet V2 family of algorithms proposed by Howard in 2018 [27] was chosen. In particular, the TF-Hub module used the TF-Slim implementation of mobilenet_v2 with a depth multiplier of 1.0 and an input size of 224x224 pixels. All images used for network training are scaled to this resolution.

This feature extractor has been trained using the ILSVRC-2012-CLS image dataset used in the google ImageNet competition. This network has been trained with an image corpus of 1.2 million images. As an unsupervised algorithm, it does not take into account the classes to which each of these images belong since the purpose of its training is to obtain a feature extractor from the image. Specifically, it uses DeepLabv3 as the feature extractor of the model proposed by Chen et al. [28] where it is explained in greater detail, but which we will briefly describe hereafter. This feature extractor uses the 3x3 Atrous convolution originally developed for the efficient computation of the undecimated wavelet transform [29] and which has been widely used for object detection [30]. A series of convolutional filters typical in these feature extraction algorithms are applied in this model, specifically the layers are based on successive copies of the blocks proposed by ResNet [31] placed in a cascade of up to 6 levels, replacing the fifth level by an Atrous Spatial Pyramid Pooling [32] with four parallel Atrous convolutions with different Atrous rates that are applied on top of the feature map because this layer has been shown to be effective to resample features at different scales allowing a more accurate classification. The module generates an output of 1280 features extracted from the original image that are then used to perform a classification.

To perform the subsequent classification, we use a pair of dense layers, the first one of 2560 neurons with ReLU activation and the

second one, a layer with N neurons, being N the number of classes that we are going to establish with SoftMax or sigmoidal activation depending on the number of classes used and the purpose of these. Further on in the specific experiments we will detail which is used in each case.

Out of the entire network, only the two dense layers were trained, leaving the convolutional feature extraction layer pre-trained with the ImageNet dataset. Next, we attempted to retrain the layer in order to detect the presence of other authors. However, the main problem found when retraining the features extractor is mainly due to the fact that the number of existing paintings by a single artist is very limited and the network lacks sufficient examples to learn correctly, even when applying augmentation techniques. Therefore, in this experiment it was decided to keep the feature extractor trained on a set of generic images and not applied to paintings. Future research would need to explore a mechanism to combine this general feature extraction provided by the module used with some other features extractor trained only on paintings, to enhance the input of dense layers with more information.

All experiments were data augmented by generating batches of 32 images with 1/255 rescaling, 50 degrees rotation, 0.25 horizontal and vertical displacement, 15 shear and zoom from 0.25 to 1.55. All experiments also used 80% of the examples for training and 20% for validation. The tests have been performed using the Keras library from Google Collaborate with GPU access.

Several experiments have been carried out using this model and variants in the dense layers, which are detailed below.

A. Rembrandt and Non-Rembrandt Detection

Taking into account the existing literature on the subject, the first step taken was to build a classifier of painting belonging to Rembrandt and those not painted by him. To maximize fake detection, the approach chosen was to sort in a binary classification between Rembrandt and non-Rembrandt paintings.

To this end, a training corpus consisting of 280 images of different resolutions was created. This was due to the fact that they were obtained from online web scrapping. All images were rescaled to a resolution of 224x224 using the OpenCV library for processing by the feature extractor.

One third of the images were paintings by Rembrandt, two thirds of the images belonged to Rembrandt's disciples and the rest were paintings by 17th century Dutch artists who influenced or were influenced by Rembrandt.

As discussed in the introduction and hypothesis, similar paintings have been selected to try to make the network generalize and learn the characteristics of the artist (Rembrandt) among similar paintings, to improve the detection of forgeries. The assumption is that the network will learn to differentiate the small subtleties of the feature vector between Rembrandt's paintings and those of his contemporaries and disciples, to then be able to generalize to AI-generated paintings. While these will present certain characteristics similar to those of Rembrandt, they should be closer to the non-Rembrandt class than to the Rembrandt class in the classification.

The goal would be to create a model capable of detecting fake images from any image generator and not only known ones. This is a basic principle of adversarial attack systems and other fake detection systems. Fake examples are used to train the network, but it is important for the system to have a good detection rate without the need to retrain the model with fake examples since models tend to overfit the data entered during training and lose their ability to generalize. It has been shown that, in the context of adversarial attacks, it is difficult to train a network with examples of attacks for learning. One will always find new examples for which the network does not behave as one expected [33]. Something similar happens in this field. If the network is trained to detect fakes produced by Dall-e 2, it does not necessarily correctly detect images generated by Imagen and vice versa. Therefore, the aim of this study is to establish specific training for already known image generation algorithms. They are an aid to the main detection algorithm, but they should not replace a more generalist model that maintains a good rate of detection of fakes, regardless of the algorithm that generates these fakes. It is because of this that the initial training corpus has not included paintings generated by these algorithms, instead it has used others that are similar but generic.

B. 17th Century Painter Detection

The Rembrandt and non-Rembrandt classification was extrapolated to 17th century artists, to include other artists. This second test has been carried out as a trial and would require further research to improve its results. This could be accomplished with a more complex network or more training examples, but results are deemed sufficient to raise within this article the possibility to extend this multiclass classification system where each painter is a specific class. A "miscellaneous" class has been used to identify other authors that are not included among the training options.

To detect 17th century Dutch painting among other artworks, two classes could be created. One would contain examples of Dutch painting from the 17th century while the second would hold examples of other authors from that century. For example, these could be non-Dutch painters. There are endless possibilities, although obviously the more complex the classifier, the machine learning model will need to be more complex, include more training examples and classification results will worsen.

As noted by Steven J. Frank, the basic problem around this domain should also be noted. Unlike image classification or object detection in images, the number of examples used in training is limited to the works produced by different painters. Rembrandt was a prolific artist with a corpus of several hundred works. Other authors do not have such a corpus available for training.

This experiment builds on the basis of the previous one. In this occasion, the network includes four classes: Rembrandt, Eeckhout, Leveck and other 17th century Dutch painters. The dataset for each class is composed of seventy sample images. Eeckhout and Leveck have been chosen due to the similarity of their style to that of Rembrandt, given that they were disciples of the renowned artist. In this experiment there is a single discard group, that of 17th century Dutch painters. After training, different tests were performed with paintings from the selected period as well as others.

The network used was based on the same feature extractor, but the classification layers are modified. A pair of dense layers are used, one of 2560 neurons with ReLU activation and the second, a layer with 4 neurons with soft max activation function.

C. Validation With Images Generated Using Dall-e 2

In the third experiment, images in the style of Rembrandt will be generated with Dall-e 2. The goal is to validate the models generated in the previous experiments in order to determine which best classifies images generated using Dall-e 2. The approach was not only to verify whether the classifier detects if a painting is a Rembrandt or not. The network output can be used to interpret the probability a processed painting has to belong to one of the categories the model classifies.

As an example, suppose we use the model from experiment 2, where there is a network that can classify 4 groups of painters. The output of the network is therefore constituted by a vector of cardinality 4. Let us imagine that once the network is trained, we expose it to a Rembrandt

painting, generating the following output [0.975, 0.750, 0.619, 0.120] with the first value being the probability of being it being painted by Rembrandt, the second the probability of being painted by Eeckhout and so on.

This output has two possible interpretations. One is to assume that the most probable class is the one that the network classifies. With this interpretation we can say that in this example the network classifies the painting as a Rembrandt. However, we can also interpret the output as a set of probabilities. Using this interpretation, we can say that the painting has a 97% chance of being a Rembrandt, a 75% chance of being an Eeckhout and a 69% chance of being a Leveck. Through this interpretation we can give more information to the network user, since with this information we can estimate the degree of confidence that the network gives to the classification. With this information we can further refine the result.

Let us imagine another scenario where the output is [0.75, 0.70, 0.40, 0.1]. In this scenario with interpretation 1 we would say that it is a Rembrandt, but if we look at the data displayed by the network, we can infer that there is a high probability of it not being a Rembrandt. The level of confidence of the network in this classification is very low for the class with the highest probability.

Three classes for the detection of fakes are proposed: Rembrandt, Fake and Doubtful.

When a painting is classified as Doubtful, the system provides a recommendation of which alternate category it leans towards, i.e., whether it is more likely to belong to the Rembrandt or Fake category. To do this the system calculates the outputs of the network in the form of probability. There is a parameter θ that determines which is the margin of doubt that best classifies the frames according to probability. To detect this parameter, we perform a study of its continuity to determine if the classification improves or worsens by applying slight changes in the parameter.

The classification function can be defined as a piecewise function as follows:

$$f(x) = \begin{cases} 2 \;\; if \quad \left(\left|c_0 - \sum_{i=1}^{N-1} C_i\right|\right) < \theta \\ 1 \;\; if \;\; C_o < \left(\sum_{i=1}^{N-1} C_i\right) \; and \; \left|c_0 - \sum_{i=1}^{N-1} C_i\right| > \theta \\ 0 \;\; if \;\; C_o > \left(\sum_{i=1}^{N-1} C_i\right) \; and \; \left|c_0 - \sum_{i=1}^{N-1} C_i\right| > \theta \end{cases} \quad (1)$$

Where C_0 is the neuron that represents the probability of a painting being a Rembrandt. C_i is the output of the network for the remaining neurons of the output layer. They indicate the probability estimated by the network of an image belonging to one of the classes predicted by the network. N is the number of classes predicted by the network.

If the result of the function is 0, it would indicate that the image is a Rembrandt. If the result is 1, it is a Fake. If the output is 2, the case is Doubtful. In this case, the system will then display the probability and also show what it predicts to be the most likely option. The expert using the system will thus have information on the reliability of the network's prediction. This is important to be able to perform a subsequent review of the doubtful cases and as we will see in the results, it improves the results obtained by the network, detecting many false positives as doubtful.

III. Results and Discussion

The results from the previously described experiments are detailed below, separated into different subsections. Results presented in this article are the best values obtained in different executions performed, as it has not been possible to minimize experiment randomness since they were executed in GPU. Due to floating point precision errors, it is extremely difficult to achieve exactly the same results in each run.

A. Results of Rembrandt and Non-Rembrandt Detection

In the first experiment performed we sought to classify the training data and its validation between two classes. These were Rembrandt and non-Rembrandt artworks. In this experiment, the hyperparameters of the dense layers have kept their default values.

The dataset was comprised of 280 images. 70 were classified as Rembrandt and 210 as non-Rembrandt. 80% were used for training and 20% for validation.

A final dense layer with sigmoidal activation was used, which is described in the literature as working very well for this type of binary classification. As a measure of loss calculation, Binary Cross-Entropy has been used, as per the following equation:

$$H_p(q) = 1 - \frac{1}{N} \sum_{i=1}^{N} y_i \dot{\log}(p(y_i)) \dot{\log}(1 - p(y_s i)) \quad (2)$$

Where y_i is the output (1 or 0) and $p(y_i)$ is the probability predicted by the network. This measure is also typical of binary classification within the literature.

The network was trained using EarlyStopping with the validation loss measure as a stopping metric.

According to these parameters, the training result produced a convergence around epoch 20 and 0.8929 accuracy.

The following step was to substitute the last layer with a SoftMax layer and change the loss metric to categorical cross entropy, as shown in Equation (3):

$$Loss = 1 - \sum_{i=1}^{N} y_i \dot{\log} p(y_i) \quad (3)$$

The output is then treated as a probability that will be used when detecting doubtful cases.

Training with these changes produced a convergence around epoch 16 with an accuracy of 0.8929 in the model validation. The results of both models are similar, but in the third experiment we will test the effectiveness of both in detecting fakes, interpreting the output as a binary classification and as a probability.

By testing the algorithm with the painting "The Next Rembrandt", a painting developed by AI, the network has yielded a 99% attribution of the painting as an authentic Rembrandt. And although a first impression would suggest that the system has failed since the painting is not really a Rembrandt, the fidelity of the result obtained by Microsoft is very high and its detection as a fake is very complex. It would probably be quite complicated to create a model that would classify it as not-Rembrandt.

B. Results of 17th Century Painter Detection

This experiment seeks to develop a classification by categories. In order to do so, the configuration from the second part of the first experiment is kept, with a last dense layer of four neurons instead of two, with SoftMax activation function. Similarly, the stopping criterion and the loss measure used are also maintained.

The results obtained from this second experiment produced a convergence around epoch 26 with an accuracy of 0.7621 in the validation of the model.

Given low image resolution and the small number of paintings used to generate the dataset, the obtained result of 0.7621, although not optimal, is acceptable for an approximation to the problem. It follows that, to optimize results, a larger dataset would be necessary, and the images used would require greater resolution. This leaves an open path for future work related to the subject through high resolution images obtained from the Rembrandt Museum in Amsterdam.

C. Results of the Validation With Images Generated Using Dall-e 2

The third experiment used the three models generated by the two prior experiments (Rembrandt and non-Rembrandt detection, 17th century painter detection) that obtained the best accuracy metrics to attempt to detect fake Rembrandt images created by the Dall-e 2 platform.

52 images were generated in the Dall-e 2 platform with the prompts "Rembrandt", "Knight painting painted by Rembrandt", "Rembrandt painted", "Rembrandt painted portrait", "Rembrandt oil portrait", and "Rembrandt-type painted portrait". Some examples of images generated by these prompts can be found in Fig. 1 and Fig. 2.

Fig. 1. Image generated by Dall-e 2 using as a prompt "Rembrandt".

Fig. 2. Image generated by Dall-e 2 using as a prompt "Knight painting painted by Rembrandt".

The two binary classifiers from the first experiment obtained an accuracy of 0.9038 and 0.8846 respectively, classifying the paintings produced by Dall-e 2 in the non-Rembrandt class.

However, as expected due to its reduced performance, the accuracy of the multi-class classifier in detecting fakes has decreased to 0.6153. To calculate this value, we considered artworks as non-Rembrandt if the algorithm attributed the painting to any of the three classes (Eeckhout, Leveck, and other 17th century Dutch painters) except Rembrandt.

These results indicate that, at least with the number of paintings in the dataset (70 per artist), the binary classifier obtains better results than a multi-class classifier when trying to detect fakes produced by Dall-e 2.

The next test used the probability generated by the second network from experiment 1 to check if the detection of fakes improved following Equation (1). In this case, the parameter θ chosen as the optimum at the authors' discretion was 0.1. This leaves us with 5% of doubtful paintings and improves the detection of fakes considering fake those classified as non-Rembrandt and doubtful to an accuracy of 0.9423. It is important not to have a high percentage of doubtful artworks since then the network would detect all paintings as doubtful. In other words, while the accuracy would be maximized, the network would not have practical sense. A value of θ that would produce more than 10% of doubtful artworks would not be useful, considering that the values obtained are already quite high.

The final experiment consisted of introducing half of the examples generated with Dall-e 2 as part of the training cases in the non-Rembrandt class. This was performed to check whether, as we supposed, the fake detection rate improved when introducing examples of the image generation algorithm in the training. However, as previously discussed in this study, caution is recommended, as the network may become overtrained when trying to detect examples generated by the algorithm as fakes and lose its ability to generalize with other unknown algorithms. Although this seems intuitively logical, it has not been tested in this work and is proposed as future research.

The results of the latter experiment converged at epoch 11 with an accuracy of 0.9655 using binary classification. In this case, the detection of dubious cases did not lead to any improvements as only one case was detected. However, it belonged to the non-Rembrandt class. As we can see in this experiment, introducing the Dall-e 2 examples in the training improves the detection of fakes obtaining better accuracy (0.9655) than the best result training without Dall-e 2 images (0.9038).

These results described in the last subsection were summarized in the Table I, where Binary classifier corresponding to the result obtained with the best model in the first experiment where the last layer had a sigmoidal activation function. The classifier Rembrandt Non-Rembrandt (R-NR) corresponds to the second part of the experiment 1, where the last layer had a SoftMax activation function.

TABLE I. Summary of the Results Obtained by the Models in the Detection of False Rembrandt With Dall-e 2

Model	Accuracy
Binary classifier	0.9038
Classifier Rembrandt – Non-Rembrandt (R-NR)	0.8846
Classifier with 4 categories	0.6153
Classifier (R-NR) witch doubtful	0.9423
Classifier (R-NR) trained with Dall-e 2 images	0.9655

IV. Conclusions

Several experiments have been performed in this study to confirm the starting hypothesis. In it, it was stated that it was possible to create a deep learning algorithm capable of detecting with a high degree of

accuracy false images generated by AI algorithms. This study focused on Dall-e 2. In this context, the best executions reached more than 90% accuracy rates.

Images generated by algorithms (Dall-e 2 in this case) were tested as a part of the training. Results improve but further work is needed to ensure that there is no loss of its generalization capacity against other algorithms that can develop this kind of forgeries or fakes. In the future, tests with other image generators such as Google Imagen or Midjourney could be performed.

The study presented has several limitations, such as those posed by the number of images available by some artists and their resolution. These elements have limited the scope of this work. This has been present in the results obtained by the second and third experiments.

Image authorship attribution presents many problems for specialists, which is why an algorithm such as the one presented in this study could potentially become a useful tool for early identification of works by an artist. However, while this would help optimize the work time of researchers, it could not replace experts when attributing authorship.

Acknowledgments

This paper was developed as part of the Quantification and prediction studies of cinematographic works (*Estudios de cuantificación y predicción de parámetros estéticos de obras cinematográficas*) research project funded by Universidad Internacional de La Rioja.

References

[1] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, "Hierarchical text-conditional image generation with clip latents," 2022, doi: 10.48550/arXiv.2204.06125.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning transferable visual models from natural language supervision," in International Conference on Machine Learning, 2021, pp. 8748–8763, PMLR.

[3] Midjourney, https://www.midjourney.com/. [Online; accessed 01-October-2022].

[4] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, et al., "Photorealistic text-to-image diffusion models with deep language understanding," arXiv preprint arXiv:2205.11487, 2022.

[5] A. Cullins, "Star Wars' and the legal issues of dead but in-demand actors," https://www.hollywoodreporter.com/movies/movie-news/carrie-fisher-star-wars-legal-issues-dead-but-demand-actors-997335/, 2017. [Online; accessed 01-October-2022].

[6] C. Santana [@DotCSV], https://twitter.com/DotCSV/status/1544959141004861441, 7 July 2022. [Online; accessed 01-October-2022].

[7] B. Edwards, "Flooded with AI-generated images, some art communities ban them completely," https://arstechnica.com/information-technology/2022/09/flooded-with-ai-generated-images-some-art-communities-ban-them-completely/, 12 September 2022. [Online; accessed 01-October-2022].

[8] Christies, https://www.christies.com/features/A-collaboration-between-two-artists-one-human-one-a-machine-9332-1.aspx, 12 December, 2018. [Online; accessed 01-October-2022].

[9] S. Lyu, "Deepfake detection: Current challenges and next steps," in 2020 IEEE international conference on multimedia & expo workshops (ICMEW), 2020, pp. 1–6, IEEE.

[10] B. Chesney, D. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," California L. Rev., vol. 107, p. 1753, 2019.

[11] R. Delfino, "Pornographic deepfakes—revenge porn's next tragic act–the case for federal criminalization," 88 Fordham L. Rev., vol. 887, 2019.

[12] H. B. Dixon Jr, "Deepfakes: More frightening than photoshop on steroids," Judges J., vol. 58, p. 35, 2019.

[13] S. Feldstein, "How Artificial Intelligence Systems Could Threaten Democracy," The Conversation, 2019.

[14] P. Rey-García, N. McGowan, La amenaza híbrida: la guerra imprevisible, ch. El deepfake como amenaza comunicativa: diagnóstico, técnica y prevención. Madrid, Spain: Ministerio de Defensa, 2020.

[15] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, "Intriguing properties of neural networks," arXiv preprint arXiv:1312.6199, 2013.

[16] S. J. Frank, "This ai can spot an art forgery." https://spectrum.ieee.org/this-ai-can-spot-an-art-forgery, 23 AUG 2021. [Online; accessed 01-October-2022].

[17] S. J. Frank, A. M. Frank, "Rembrandts and robots: Using neural networks to explore authorship in painting," arXiv preprint arXiv:2002.05107, 2020.

[18] W. Zhao, D. Zhou, X. Qiu, W. Jiang, "Compare the performance of the models in art classification," Plos one, vol. 16, no. 3, p. e0248414, 2021. https://doi.org/10.1371/journal.pone.0248414

[19] S. J. Frank, A. M. Frank, "Salient slices: Improved neural network training and performance with image entropy," Neural Computation, vol. 32, no. 6, pp. 1222–1237, 2020. https://doi.org/10.1162/neco_a_01282

[20] Y. Wu, Q. Wu, N. Dey, S. Sherratt, "Learning Models for Semantic Classification of Insufficient Plantar Pressure Images," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 1, pp. 51-61, 2020. https://doi.org/10.9781/ijimai.2020.02.005

[21] F. Xu, T. Wu, S. Huang, K. Han, W. Lin, S. Wu, S. CB, S. R. Dinesh Jackson, "Extensive Classification of Visual Art Paintings for Enhancing Education System using Hybrid SVM-ANN with Sparse Metric Learning based on Kernel Regression," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 7, no. 2, pp. 224-231, 2021. https://doi.org/10.9781/ijimai.2021.10.001

[22] A. Blessing, & K. Wen, Using machine learning for identification of art paintings. Technical report. 2010.

[23] A. Oliva, A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," International journal of computer vision, vol. 42, no. 3, pp. 145–175, 2001.

[24] N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection," in 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol. 1, 2005, pp. 886–893, IEEE.

[25] S. Lazebnik, C. Schmid, J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06), vol. 2, 2006, pp. 2169–2178, IEEE.

[26] M. J. G. Narag, M. N. Soriano, "Identifying the painter using texture features and machine learning algorithms," in Proceedings of the 3rd International Conference on Cryptography, Security and Privacy, 2019, pp. 201–205.

[27] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.C. Chen, "Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation," arXiv preprint https://arxiv.org/abs/1801.04381v2, 2018.

[28] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, "Rethinking atrous convolution for semantic image segmentation," arXiv preprint arXiv:1706.05587, 2017.

[29] M. Holschneider, R. Kronland-Martinet, J. Morlet, P. Tchamitchian, "A real-time algorithm for signal analysis with the help of the wavelet transform," in Wavelets, Springer, 1990, pp. 286–297.

[30] J. Dai, Y. Li, K. He, J. Sun, R. Fcn, "Object detection via region-based fully convolutional networks," arXiv preprint arXiv:1605.06409, 2016.

[31] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[32] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 4, pp. 834– 848, 2017.

[33] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, P. Frossard, "Universal adversarial perturbations," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1765–1773, 2017.

Marcelo Fraile-Narváez

Marcelo Fraile-Narváez received a PhD in architecture from the University of Buenos Aires. He is a specialist in new technologies and biodigital design. He has taught undergraduate and postgraduate courses at different institutions including the University of Costa Rica, the International University of Catalonia, the Complutense University of Madrid, and the King Juan Carlos University. He is currently a lecturer at the School of Engineering and Technology of the International University of La Rioja (UNIR) where he is also the Academic Coordinator of the master's degree in Multimedia Design and Production. His research interests include the use of artificial intelligence methods and biology to aid the architectural design process.

Ismael Sagredo-Olivenza

Ismael Sagredo-Olivenza received a Ph.D. degree from the Complutense University in Madrid for his research in artificial intelligence applied to video game design and development. He worked as a Programmer in the video game industry in studies, such as Pyro Studios and Padaone Games. He is a currently a Professor with the High School of Engineering and Technology, International University of La Rioja (UNIR). He is also the Director of the MSC Program in videogames design and development with UNIR. In the last one, he developed some serious and educational games. His research interest includes use of artificial intelligence methods to help the video game design process.

Nadia McGowan

Nadia McGowan holds a degree in Cinematography (ECAM), Bachelor's in Art History (UNED), Master's in Screenwriting (UNIR), and Doctorate Audiovisual communication, advertising, and public relations (Universidad Complutense). She has worked at Notre Dame University and the Lebanese German University in Lebanon and currently is part of the Design Department of the School of Engineering and Technology at Universidad Internacional de La Rioja. Her research is focused on technical aspects of filmmaking.

Detection of Improperly Worn Face Masks using Deep Learning – A Preventive Measure Against the Spread of COVID-19

Anubha Bhaik, Vaishnavi Singh, Ekta Gandotra*, Deepak Gupta

Department of Computer Science and Engineering, Jaypee University of Information Technology, Waknaghat, Solan, HP (India)

Received 10 November 2020 | Accepted 24 June 2021 | Published 9 September 2021

Abstract

Coronavirus disease 2019 has had a pressing impact on people all around the world. Ceasing the spread of this infectious disease is the urgent need of the hour. A vital method of protection against the virus is wearing masks in public areas. Not merely wearing masks but wearing masks properly can ensure that the respiratory droplets do not get transmitted to other people. In this paper, we have proposed a deep learning-based model, which can be used to detect people who are not wearing their face masks properly. A convolutional neural network model based on the concept of transfer learning is trained on a self-made dataset of images and implemented with light-weighted neural network called MobileNetV2 for mobile architectures. OpenCV is used with Caffe framework to detect faces in an input frame which are further forwarded to our trained convolutional neural network for classification. The method has been implemented on various input images and classification results have been obtained for the same. The experimental results show that the proposed model achieves a testing accuracy and training accuracy of 93.58% and 92.27% respectively. Optimal results with high confidence scores and correct classification have also been achieved when the proposed model was tested on individual input images.

Keywords

COVID-19, Mask Classification, MobileNetV2, OpenCV, Transfer Learning.

DOI: 10.9781/ijimai.2021.09.003

I. Introduction

THE coronavirus disease 2019 (COVID-19) is an infectious disease that can result in mild to severe illnesses in people infected by it. It is transmitted mainly through respiratory droplets of saliva or discharge from the nose when a person infected with coronavirus coughs or sneezes. Thus, it is essential to practice a proper respiratory protocol such as covering our face while sneezing or coughing [1]. The mucous membranes of the face should be covered properly with protective equipment to protect oneself and others from the continual transfer of the disease.

Face masks are being used by people all over the world now. In many countries, it is now compulsory to wear a face mask when stepping out of home. However, many people do not wear face masks properly. They fidget with their masks and pull them under their noses or completely off their faces to rest under their chins without realizing that improperly wearing a mask leads to an increased risk of contamination. Wearing a face mask limits the spread of the virus from someone who knows or does not know they have an infection or not. It also reminds others to continue the practice of physical distancing [2]. Moreover, the presence of asymptomatic spreaders of the COVID-19 virus means that wearing a face mask should be a part of lessening the cases of COVID-19 [3]. Thus, the masks must be worn properly covering the mouth and nose appropriately to prevent respiratory droplets from spreading. A recent study [4] has pointed out that surgical face masks can prevent the transmission of coronavirus and influenza viruses from symptomatic people. Another study [5] states that the reproductive rate of COVID-19 is higher as compared to the SARS coronavirus, and thus it is essential to wear masks properly as a measure to keep public health in mind. Countries have been exiting the lockdown lately to reduce the effect of the pandemic on the economy, but the coronavirus persists as an inevitable danger in most of the countries. Since the outbreak of this disease is not only the concern of a single country but of the entire world. The stringent measures implemented by the government have been effective in combating the spread of COVID-19 disease [6]. For instance, China has been using mass surveillance to monitor people and track the spread of coronavirus. Other nations are also deploying technologies like video camera footage, credit card information, and location tracking as they race against the outbreak. Surveillance of activities of people can be effective as we can monitor whether people are properly taking protective measures, and by not letting them enter a public place if they are careless about protection. A recent study [7] has established that face masks have been effective in the containment of COVID-19 in South Korea. This study further states that in addition to maintaining social distancing and sanitizing hands, properly wearing appropriate masks has been efficacious in lessening severe cases in South Korea. Another study [8] focuses on how the universal

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

use of covering the face by even a simple cloth mask, if not a surgical mask, can help in acting as a preventive measure.

With this study, we would contribute to public healthcare by detection of people not wearing their masks properly at places where the chances of getting infected are high. People present at places where there are low chances of enforcing social distancing, can be checked for wearing their masks properly, especially in severely affected zones of countries. We intend to use mobile devices to check whether someone is wearing their masks properly or not. The main contributions of the paper are:

1. A dataset containing images of people wearing masks properly and improperly.
2. A model for the detection and classification of faces wearing masks properly and improperly.
3. Experiments to evaluate the performance of the proposed model on a dataset using various evaluation metrics.

This paper is organized as follows: Section II discusses the background and the work related to our study. Section III describes the detailed methodology used for the proposed model. Further, Section IV presents the analysis and visualization of the experimental results followed by the conclusion in Section V.

II. Related Work

This section discusses the related research work behind the proposed model for detecting improperly worn face masks.

A. Convolutional Neural Networks

Convolutional neural network (CNN) is a class of deep learning models that largely deals with the analysis of visually descriptive data. CNNs can extract important features from visual data without human supervision with the help of various layers. Different layers perform different kinds of transformations on the data. CNNs treat data as spatial and can simplify the complexity of images to be better understood and processed by the machine and hence are widely used for pattern recognition. A classic CNN is composed of multiple layers namely convolutional layers and pooling layers which are used for the extraction of important features from input data. It also has some layers in the end which take the output from the two mentioned layers and help in classifying the data into labels. CNNs have wide applications like face detection and recognition, classification of malware applications [9], classification of X-ray images [10], etc. Image classification is one of the most popular applications of CNN. Sultana et al. [11] have done a study where they explained different architectures of CNN used for image classification. Shinet et al. [12] explored and evaluated different CNN architectures and discussed when and why transfer learning from pre-trained ImageNet CNN models can be valuable. Demir et al. [13] extracted distinctive face features using CNN and used the Softmax classifier to classify faces in the fully connected layer of CNN. In [14], the inception network has been proposed for allowing the network to learn the best combination of kernels, leading to an effective image classification method as well. Wang et al. [15] proposed the residual attenuation network for image classification achieving 0.6% top-1 accuracy improvement as compared to ResNet-200. They focused on the depth of the network and used the attention mask mechanism to take image classification to a new level.

B. Object Detection

Object detection is a computer vision approach used to identify the objects present in images or videos. Many deep learning-based frameworks for object detection have been proposed in the literature, covering various aspects of its applications in the real world like facial recognition, face detection, detection of obstacles for self-driving cars, and more. A review on some object detection architectures has been carried out by Zhao et al. [16]. Other fast techniques like You Only Look Once (YOLO) [17] have also been proposed for object detection.

C. Face Detection

Face detection is a popular application of object detection that is being widely used today. A comprehensive survey of various techniques for facial detection in digital images is due to Kumar et al. [18]. Sivaram et al. [19] proposed a technique that uses recurrent neural networks (RNN) and deep neural networks (DNN) to take in the shape of the face for accurate facial detection. A camera-based PCA facial recognition system has been built by Khan et al. [20] using programming on technologies like OpenCV, Haar Cascade, and Python.

Face detection or recognition systems have been in demand for several security-based applications too, like surveillance or tracking of suspected people, access management, etc. Zhang et al. proposed a framework for serving better surveillance functionality for ATMs. The technique included tackling severely occluded faces by fusing the features of faces like skin color and facial structure; it achieved 98.56% accuracy on detection of face occlusion [23].

D. Face Mask Detection

Face mask detection has also been explored by researchers to tackle the situation of COVID-19 for ensuring if people are wearing masks or not. In [21], Meenpal et al. have studied facial mask detection using semantic segmentation. They have proposed a binary face classifier that can detect any face in the frame. Their method uses pre-trained weights of VGG-16 architecture for feature extraction and the experimental results give a mean pixel-level accuracy of 93.88% for the segmented face masks. Besides, Jiang et al. [22] have proposed a face mask detector that is able to detect face masks. They have tried to distinguish between people wearing masks and people merely covering their mouths with their hands.

This pandemic certainly demands the need for proper mask detection for the security of the health of citizens. However, in the aforementioned studies, the authors have not considered whether a person is wearing a face mask properly. The algorithms proposed in these studies only detect if the person is wearing a face mask or not. In this paper, we have proposed a CNN-based model that addresses this task of checking if the person is wearing the mask properly.

III. Proposed Methodology

This section describes the proposed methodology for the detection of improperly worn masks. Fig. 1 shows the complete workflow of the used methodology.

A dataset consisting of images of people wearing masks both properly and improperly is created and used in this work. These images were collected from the local, existing datasets, and the Internet. The images were pre-processed to enhance their generalization during the training of the CNN. Further, CNN is trained on the created dataset for classification purpose. For visualization of results, OpenCV and Caffe framework are used. The model creates a bounding box around a person's face in the input image classifying whether the mask worn is proper or improper. Finally, the experimental results are analyzed and visualized. The various steps of the proposed methodology are further elaborated in the following sections.

A. Data Acquisition

The images of people wearing proper face masks are collected from the images present in existing datasets [24] and various other Internet sources. Since these datasets had fewer images of improperly

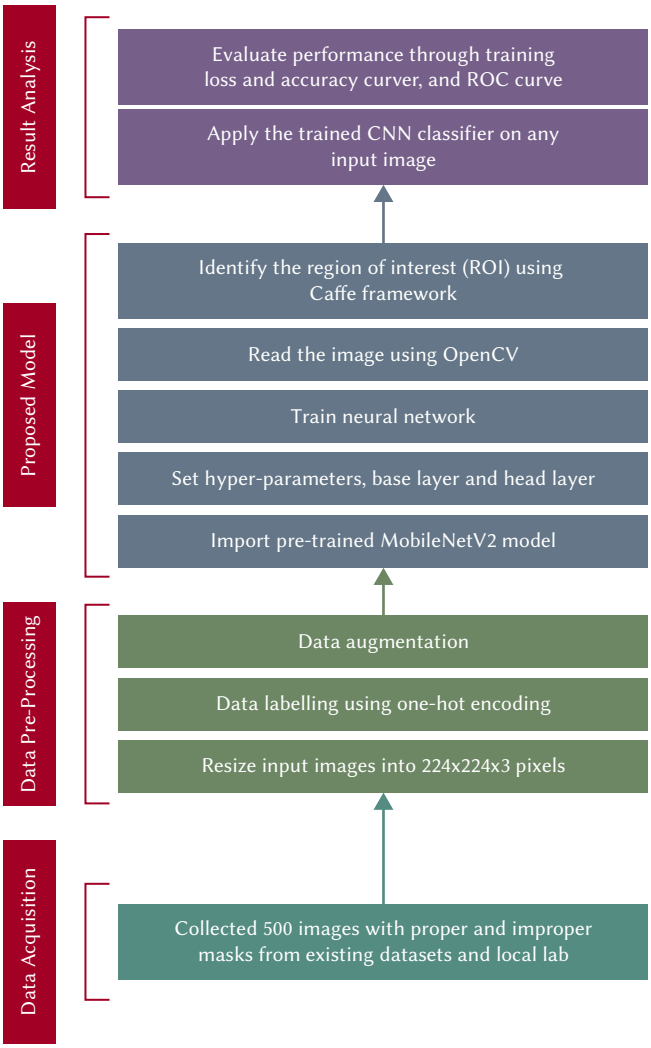


Fig. 1. Workflow of the methodology used.

worn masks, we collected such images from the Internet and local lab. Finally, our dataset consists of 500 images, equally distributed among properly and improperly worn mask categories. Figs. 2 and 3 show the sample images from the dataset of people with proper and improper masks, respectively.

Fig. 2. Sample images of the dataset belonging to the class – proper mask.

Fig. 3. Sample images of the dataset belonging to class – improper mask.

B. Data Pre-processing

The dataset consists of images of different sizes, and thus, these images are converted into a uniform size of 224 × 224 pixels. After the application of RGB reshaping, a 224 × 224 × 3 image is given as input to the proposed model. The class labels are one-hot encoded. These pre-processed images and encoded labels are added to separate lists, one for the pre-processed images and the other for the class labels. Furthermore, data augmentation strategies like random rotation, shift, shear, zoom and flip, are applied on the images for increasing the generalization of data which help in improving the performance of the model.

C. Proposed Model

This section describes the proposed model. The neural network for classification is built and trained after setting the various hyper-parameters and hidden layers. Thereafter, the trained classifier is applied to input image for classification into 'proper' and 'improper'.

1. Training the Neural Network

The problem for our proposed model is to learn the interpretation of various features in images and classify them accordingly. CNNs help in leveraging the spatial information in images. Fig.4 presents a basic architecture of CNN which consists of the input images, the layers of the network, and the corresponding output.

We have split our dataset into training set and testing set in such a way that 80% data is used for the training purpose and 20% data is used for the testing purpose. For achieving the optimum results on our dataset, we use an aspect of deep learning called transfer learning. Transfer learning is the act of transferring the knowledge previously gained by one model on a specific task to a new similar task that will benefit from some or all the layers of the previously built model. To aid the use of our model on mobile devices, we used MobileNetV2 as the base model. It has less computation cost and is an efficient mobile-oriented model for transfer learning [25]. Additionally, it is an effective feature extractor used for object detection that improves the performance of our detection. The pre-trained weights for the ImageNet [26] dataset have been used as the backbone.

Unlike the typical convolution, MobileNetV2 utilizes an advanced version of convolutional operation called the depth-wise separable convolution which leads to a lesser number of computations and transformations on the images than the conventional convolution. It gets applied to images in two parts [27]. The first part is the depth-wise convolution used to perform the filtering stage and the second part is pointwise convolution for the combining stage. It is a light-

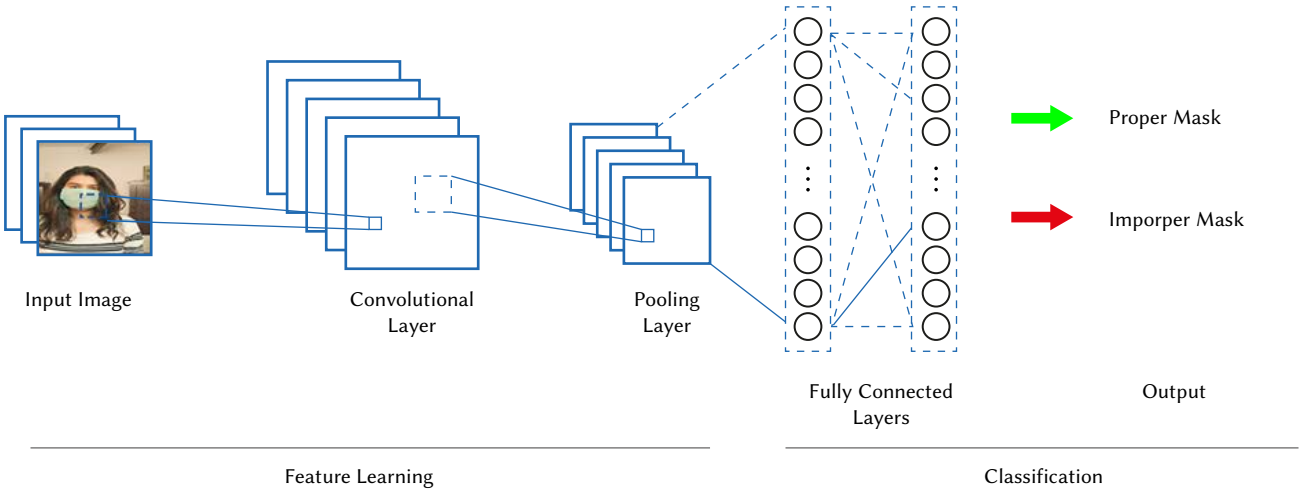


Fig. 4. Elementary CNN architecture for image classification.

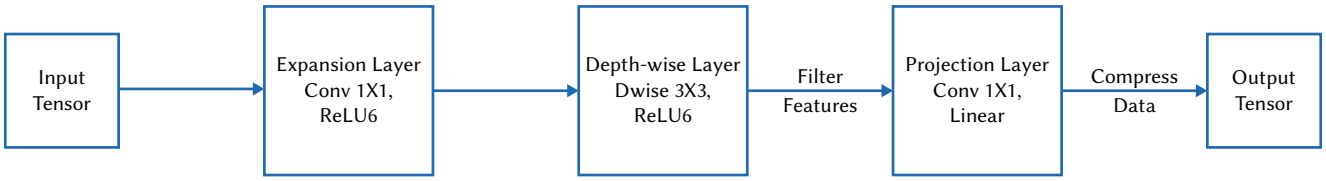


Fig. 5. Basic architecture of a block in MobileNetV2.

weight model with low-latency which provides comparable accuracy to other heavy and complicated models. Since the size of our dataset is small and its affinity to ImageNet is not that high, we fine-tune the top layers of MobileNetV2 for our work. The head layers of the pre-trained MobileNetV2 CNN architecture are unfrozen while importing it and are replaced with new custom layers. The weights for the input to this pre-trained MobileNetV2 architecture are set by default. All base layers (we call these as BaseModel) below the head of the MobileNetV2 architecture are frozen to prevent their weights from getting updated during backpropagation. The first layer in the network is a fully convolutional layer with input size $224 \times 224 \times 3$.

The fundamental building block of MobileNetV2 architecture is a bottleneck depth-separable convolution with residuals, so the input layer is followed by residual bottleneck layers [27]. Each block of the model consists of 3 convolutional layers as shown in Fig.5. First is a 2D convolutional layer (expansion layer) which performs 1×1 convolution for expansion of the number of channels in the data, then batch normalization and ReLU6 non-linearity are applied, which limits the maximum value of activation to 6. The default expansion factor is taken as 6 in the expansion layers. The second is a depth-wise convolution layer, again with batch normalization and ReLU6 non-linearity. Some of the convolutional layers have a stride of 2 for achieving spatial down-sampling since there are no conventional pooling layers, other layers are kept at a stride of 1. The third layer known as a pointwise convolutional layer (projection layer) performs linear convolution to reduce the dimensionality of input (also known as a bottleneck layer) and again accompanied by batch normalization [27]. The first block of the model is different as it comprises 3×3 convolution with 32 channels rather than the default 1×1 convolution which happens in the expansion layer.

The last five custom layers (we call these as HeadModel) which produce output for the model include the average pooling 2D layer with pool size 7×7, reducing the dimensionality by acquiring average values from each region of the image. This layer precedes a flatten layer that reshapes the pooled feature map to a single column vector. The simple feature vector is now put into a dense layer of 128 units of size accompanied by ReLU activation by using (1).

$$f(y) = \{y \; for \; y \geq 0, 0 \; for \; y < 0\} \tag{1}$$

A dropout layer is applied on this dense layer to prevent the model from overfitting, with a threshold value of 0.5. Then a final dense layer is applied with Softmax non-linear activation by using (2) to provide two output values, i.e., probabilities of the image belonging to the proper and improper mask groups, respectively.

$$\sigma(x_j) = \frac{e^{x_j}}{\sum_i^J e^{x_i}} \tag{2}$$

where x is a vector of the inputs in the form of images to the output layer and i indexes the output unit such that $i - 1, 2, 3, \ldots f$. The detailed summary of our proposed model consisting of the MobileNetV2 model (BaseModel) and the custom layers (HeadModel) is provided in Appendix A.

We use Adam optimizer for the optimization of the CNN and binary cross-entropy as loss function as shown in (3). This loss function is used in a binary classification problem.

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i . log(p(y_i)) + (1 - y) . log(1 - p(y_i)) \tag{3}$$

Here, y is the label (which has been one-hot encoded) and p(y) is the predicted probability of one of the labels.

The initial learning rate is set to 0.0001. A learning rate decay schedule (represented by (4)) is created which helps in increasing the model accuracy and descend into areas of lower loss. Table I shows hyperparameters used in our CNN Model.

$$\alpha = \frac{1}{1 + decay * iteration} * \alpha_0 \tag{4}$$

Where α and α_0 represent the learning rate and initial learning rate respectively, and

$decay = \alpha_0 / total \; number \; of \; epochs$

TABLE I. Hyperparameters Used in Our CNN Model

Parameter Name	Value
Kernel Size	3 x 3
Activation Function	ReLU6 (BaseModel), ReLU (HeadModel)
Average Pooling	7 x 7 (HeadModel)
Optimizer	Adam
Loss Function	Binary cross entropy
Dropout	0.5
Epochs	50
Batch Size	32
Initial Learning Rate	0.0001
Fully Connected (Activation Function)	Softmax

2. Detection of Region of Interest (ROI)

After training the CNN, we have used a deep learning framework, Caffe, along with an open-source computer vision library (OpenCV) for face detection using static input images. For extracting the region of interest (ROI) in the image, the DNN module of OpenCV is used with Caffe. The network model which is stored in Caffe framework format (with the learned network) and a file containing the text description of the network architecture is read using OpenCV DNN module. The file with the Caffe framework format has been provided by the OpenCV for face detection [28], [29] and contains the weights for the actual layers. The Caffe model is based on the Single Shot MultiBox Detector (SSD) framework which uses ResNet as a base network for facial recognition [30].

The trained model is used on various static input images to detect whether the person is wearing the mask properly over his nose. An input image is first uploaded and pre-processed using OpenCV DNN module. The spatial dimensions of the input image are extracted and converted into a 4D Binary Large OBject (BLOB) which is further used to perform functions like scaling, mean subtraction, and resizing on the input image. The scale factor is set to 1.0. After normalizing the input image to create a BLOB, it is passed through the DNN to obtain face detections. The detections obtained are further checked for the probability or confidence which is used to classify the input image as proper or improper. The threshold confidence (or probability) is kept at 0.5 to filter all the weak detections. Further, OpenCV is used to extract the region of interest (ROI) of the face which helps in displaying the bounding box. The extracted face ROI is converted from BGR to RGB ordering of channels and the image size is set to 224 × 224 pixels to pass it through the trained model. Finally, this pre-processed input image is passed through the trained model to determine if the mask is worn correctly or not. This can finally be visualized by a bounding box labelled with the class score in the image. The class score is the probability that the image contains a face with a proper or an improper mask.

3. Algorithm

The proposed algorithm (as shown in Algorithm I) is based on transfer learning and facial detection using OpenCV. θ is the initial learning rate which is used to update weights during the training phase. It has a value that is often in the range of 0.0 to 1.0. β denotes batch size which refers to the number of training images that are used in a single iteration. ϕ represents epochs which define the total number of times the model will iterate over the training images. The values of these three hyper-parameters (θ, β and ϕ) can be tuned based on Steps 3-5 in a hit and trial way to achieve better accuracy results. The updated output weights (ω) are stored in a .h5 file which are used further to get predictions on other random input images. δ gives the number of face detections that are obtained on our static input image. α is used to filter out all the weak face detections in our input image and is known as the confidence (or probability) for detection of facial features.

The algorithm advances by using the architecture of MobileNetV2 as the base layers (refer to Step 4) and the addition of custom head layers (refer to Step5). A for loop is applied in the range of the total number of defined epochs which updates the weights of only the custom layers through forward and back propagation (refer to Steps 8-10). The input static image is uploaded, and features extracted for facial detection (refer to Steps 11-13). Another for loop is applied to detect the region of interest in the input image and plot a bounding box with the indicative predicted probability as a label (refer to Steps 13-16). Finally, the trained model is applied to the input image only if facial features are detected (refer to Step 17).

Algorithm I. Algorithm for Detection of People Wearing Proper and Improper Face Masks

Input: Images of size h×w×d; where h⟶height of image,w⟶width of image,d⟶number of channels in the RGB image.

$\theta \longrightarrow$ initial learning rate

$\beta \longrightarrow$ number of samples of images trained in one iteration

$\phi \longrightarrow$ number of epochs

$\delta \longrightarrow$ number of face detections on input image

$\alpha \longrightarrow$ numeric constant to filter out weak detections

Output: classified output image with probability of prediction

Begin:

1. Split the input images randomly into training set ($\sigma 1$) and testing set ($\sigma 2$) using 80% data for the training set and the remaining 20% for the test set.
2. Construct the image generator for augmentation of images.
3. Initialize the CNN parameters θ, $\beta$and ϕ.
4. Determine the base layers of CNN architecture, i.e., MobileNetV2.
5. Set the head layers, $CNN_{averagepooling2D}$, $CNN_{flatten}$, CNN_{dense}, $CNN_{dropout}$
6. Set the last layer (at the end of the fully connected layers) that contains labels for classification.
7. Train the CNN to compute ω.

for every ϕ:

8. Select a mini batch of size β from $\sigma 1$.
9. Forward propagation and compute loss (binary cross-entropy) via θ.
10. Back propagation only on the head layers, update ω with Adam optimizer.

end

11. Upload static input image and convert it into a 4-D Binary Large OBject (BLOB).
12. Initialize α (confidence).
13. Set the Caffe framework for obtaining face detections.
14. Pass BLOB image through the network to obtain face detections on the input image.

for every δ :

15. Extract probability (confidence) associated with the detection.
16. Filter weak detections by comparison with α
17. Extract face ROI in the input image, add bounding box and labels on face.

end

18. Apply the trained model to the image only if a face is detected.

End.

IV. Evaluation

This section describes the evaluation parameters and the outcomes of the proposed solution. The experimental results are further analyzed and visualized using the performance evaluation metrics.

A. Performance Parameters

The description of various evaluation parameters is given as follows:

- **Confusion Matrix**: A confusion matrix is an n-way matrix where the n is the number of classes for the classification purpose. It is based on four important parameters: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The following parameters are calculated using the confusion matrix:

  Accuracy: It is calculated by using (5).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

  Precision: It is another evaluation metric that tells how many predictions are actually correct out of all the correct predictions. It is given by (6).

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

  Recall: defined as the number of positive predictions made by the model out of the total actual positive classes. It is calculated by (7).

$$Recall = \frac{TP}{TP+FN} \tag{7}$$

  F1-score: It is calculated by using precision and recall as shown in (8).

$$F1-score = \frac{2}{\frac{1}{Precision}+\frac{1}{Recall}} \tag{8}$$

  Sensitivity: It measures how correctly we have detected the positive classes. It is calculated by using (9).

$$Sensitivity = \frac{TP}{TP+FN} \tag{9}$$

  Specificity: It measures how exact or accurate is the assignment to the positive class. It is calculated by using (10).

$$Specificity = \frac{TN}{FP+TN} \tag{10}$$

- **Learning Curves**: A learning curve is a plot for the graphical visualization of the model performance while it is training on the dataset. It is a plot of the accuracy/loss versus the number of epochs. It shows how the accuracy/loss of the model changes during its training phase with the increase in the number of epochs.
- **Receiver Operating Characteristics**: The ROC represents a plot between the true positive rate (TPR) and false negative rate (FNR), and is a trade-off between the specificity and sensitivity [31]. AUC uses the ROC curve and is calculated by using the trapezoidal method, i.e., dividing the area into a number of sections with equal width. Here, the trapezoid (T) refers to integration of points (a, b) from a functional form which is divided into n equal pieces. The addition of the area of each section of the trapezium formed when the upper end is replaced by a chord and the sum of these approximations provides the final AUC value. The trapezoidal formula is indicated as an integral of the function $\int_a^b f(x)\,dx$, and the points of integration (a, b) are labelled as $\{x_0, x_1, ..., x_n\}$; where $\{x_0 = a, x_n = b, x_r = x_0 + r(b - a)/n\}$ as given in [32].

B. Evaluation Results

In our evaluation, we noted the model performance in terms of accuracy, precision, recall, F-score, sensitivity, specificity, and area under curve (AUC) of receiver operating characteristics (ROC). The training history of the model is also plotted for a credible analysis of the loss and accuracy of the training set and the validation set. Confusion matrix is calculated on the test set as shown in Table II.

TABLE II. Confusion Matrix Obtained on Test Data

		Predicted Classes	
		Proper	Improper
Actual Classes	Proper	TP = 47	FN = 3
	Improper	FP = 4	TN = 55

By calculating the overall accuracy of the proposed model by using the confusion matrix on the testing data, we have attained an overall accuracy of 93.58%. We have achieved a precision and recall of 92.15% and 94% respectively. Further, we have achieved F-1 score of 93.10. A sensitivity of 94% has been calculated. This means that out of 50 total people who are wearing the masks properly, 47 have been detected wearing the mask properly from the test set.

We have achieved a specificity of 93.22%. This implies that out of 59 people who are wearing the masks improperly, we are able to correctly predict the person wearing an improper mask with an error rate of only 6.78%. These evaluation parameters are calculated from the confusion matrix as indicated in Table III.

TABLE III. Summary Table of Computed Metrics Using the Confusion Matrix

	Evaluation Metric	Value (in %)
1.	Accuracy	93.58
2.	Precision	92.15
3.	Recall	94.00
4.	F-1 score	93.10
5.	Sensitivity	94.00
6.	Specificity	93.22

The model has been trained for 50 epochs, with an initial learning rate of 0.0001. After the proposed model is trained on the training set data, we observed the training accuracy as 92.27% and the validation accuracy is 93.58% as shown in Fig. 6. The validation set data is used to provide an unbiased evaluation and tune the hyper-parameters of the model, while the model is fit on the training set data. It further helps in determining the error rate in the model by holding out a subset of the data from the fitting process and evaluation of the loss of the model at the end of each epoch. The high training accuracy can be considered as a good measure to assess our classification model. Fig. 7 depicts learning curves which gives the values of training and validation loss as 0.1693 and 0.1595, respectively.

Fig. 6. Learning curves for evaluation of Accuracy.

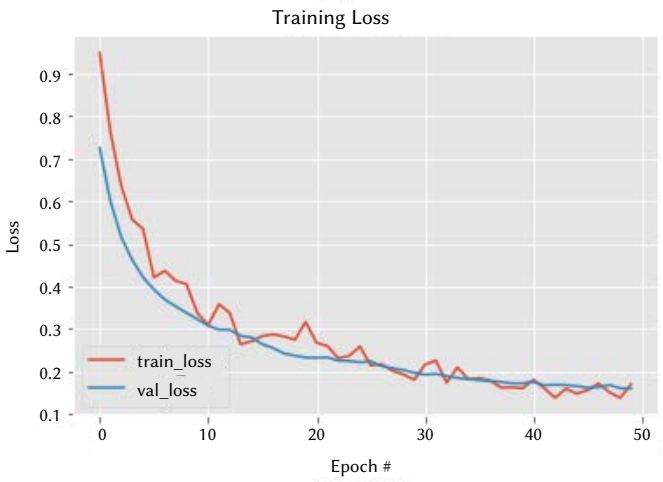


Fig. 7. Learning curves for evaluation of Loss.

Table IV depicts the relevant results obtained from the training curves to determine and compare the loss and accuracy of the training and validation data, respectively. Fig. 8 shows the plotted AUC of ROC curve of the proposed model which is above the threshold level and is calculated as 0.9361.

TABLE IV. Inference of Accuracy and Loss From Obtained Curves and Their Training Vs Validation Comparison

	Accuracy/Loss	Value (in %)
1.	Training Accuracy	92.27
2.	Validation Accuracy	93.58
3.	Training Loss	0.1693
4.	Validation Loss	0.1595

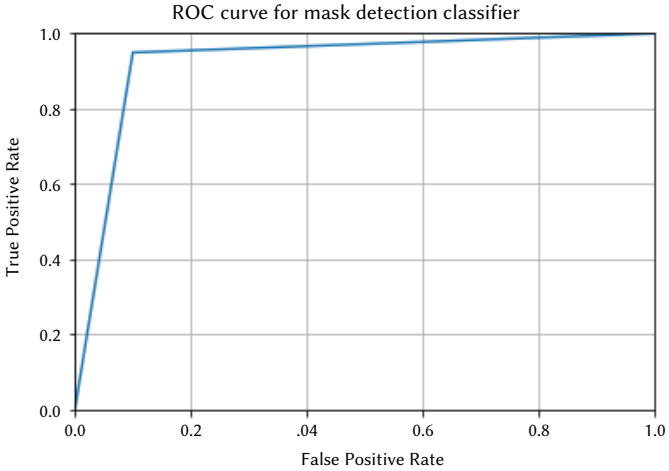


Fig. 8. ROC curve for test data.

The images that were not used in training are provided as input to the proposed model to predict whether they are wearing the mask properly. Fig. 9 depicts that the proposed model is able to categorize faces into two classes: proper and improper, with high confidence scores. Clockwise from top left, confidence scores are 99.06% (Improper), 96.23% (Proper), 59.90% (Improper), 74.18% (Proper), and 99.81% (Proper).

We have not compared our results to any existing work since these do not focus on whether the person is wearing a face mask properly or not.

Fig. 9. A sample of results showing successful classification achieved by the model on input images.

V. Conclusion

The COVID-19 disease is the greatest challenge that the world has faced since World War II. To prevent its rapid spread, face masks must be worn properly by people over their noses. People at crowded places, hospitals, offices, and working spaces can be checked for improperly worn masks to ensure safety. Application of the proposed model can serve as a preventive measure in the COVID- 19 crisis and benefit in safekeeping the health of society. The government can also leverage the model to mobile devices or any device with low computational power for the detection of improperly worn face masks at public places.

Some research work had been done in detecting masks worn and not worn. However, our model specifically focused on classifying the mask worn by a person into two classes: proper and improper. This will be much significant in the various stages of unlocking all over the world as it will contribute to public safety and healthcare. The architecture of the model consists of the light weighted MobileNetV2 neural network as the backbone which overcomes computational issues as it can be used efficiently on devices with low computational power. Transfer learning has been adopted to use weights that have been used for a similar task like face detection and already trained on a very large dataset. Furthermore, OpenCV with the Caffe framework has been used to detect facial features on experimental input images and used on the pre-trained model with our dataset, to produce classification results with indicative results, such as labels and a bounding box. We are able to attain a testing accuracy of 93.58% and an AUC measure of 0.936.

Appendix

The appendix summarizes the complete structure of the proposed model. It includes information about the layers and their order in the model, the output shape of each layer and the information about the parameters (weights). The number of parameters in each layer and the total number of parameters in the model are obtained by using the model summary. A total of 2,422,210 parameters (164,226 trainable parameters and 2,257,984 non-trainable parameters) is present in our model.

Layer(type)	Output Shape	Param #
input_1 (InputLayer)	[(None, 224, 224, 3)	0
Conv1_pad (ZeroPadding2D)	(None, 225, 225, 3)	0
Conv1 (Conv2D)	(None, 112, 112, 32)	864
bn_Conv1 (BatchNormalization)	(None, 112, 112, 32)	128
Conv1_relu (ReLU)	(None, 112, 112, 32)	0
expanded_conv_depthwise (DepthwiseConvolution)	(None, 112, 112, 32)	288
expanded_conv_depthwise_BN (BatchNormalization)	(None, 112, 112, 32)	128
expanded_conv_depthwise_relu (ReLU)	(None, 112, 112, 32)	0
expanded_conv_project (Conv2D)	(None, 112, 112, 16)	512
expanded_conv_project_BN (BatchNormalization)	(None, 112, 112, 16)	64
block_1_expand (Conv2D)	(None, 112, 112, 96)	1536
block_1_expand_BN (BatchNormalization)	(None, 112, 112, 96)	384
block_1_expand_relu (ReLU)	(None, 112, 112, 96)	0
block_1_pad (ZeroPadding2D)	(None, 113, 113, 96)	0
block_1_depthwise (DepthwiseConvolution)	(None, 56, 56, 96)	864
block_1_depthwise_BN (BatchNorm	(None, 56, 56, 96)	384
block_1_depthwise_relu (ReLU)	(None, 56, 56, 96)	0
block_1_project (Conv2D)	(None, 56, 56, 24)	2304
block_1_project_BN (BatchNormalization)	(None, 56, 56, 24)	96
block_2_expand (Conv2D)	(None, 56, 56, 144)	3456
block_2_expand_BN (BatchNormalization)	(None, 56, 56, 144)	576
block_2_expand_relu (ReLU)	(None, 56, 56, 144)	0
block_2_depthwise (DepthwiseConvolution)	(None, 56, 56, 144)	3456
block_2_depthwise_BN (BatchNormalization)	(None, 56, 56, 144)	96
block_2_depthwise_relu (ReLU)	(None, 56, 56, 144)	0
block_2_project (Conv2D)	(None, 56, 56, 24)	3456
block_2_project_BN (BatchNormalization)	(None, 56, 56, 24)	576
block_2_add (Add)	(None, 56, 56, 24)	0
block_3_expand	(ne, 56, 56, 144)	0
block_3_expand_BN (BatchNormalization)	(None, 56, 56, 144)	1296
block_3_expand_relu (ReLU)	(None, 56, 56, 144)	576
block_3_pad (ZeroPadding2D)	(None, 57, 57, 144)	0
block_3_depthwise (DepthwiseConvolution)	(None, 28, 28, 144)	4608
block_3_depthwise_BN (BatchNormalization)	(None, 28, 28, 144)	128
block_3_depthwise_relu (ReLU)	(None, 28, 28, 144)	6144
block_3_project (Conv2D)	(None, 28, 28, 32)	768
block_3_project_BN (BatchNormalization)	(None, 28, 28, 32)	0
block_4_expand (Conv2D)	(None, 28, 28, 192)	1728
block_4_expand_BN (BatchNormalization)	(None, 28, 28, 192)	768
block_4_expand_relu (ReLU)	(None, 28, 28, 192)	0
block_4_depthwise (DepthwiseConvolution)	(None, 28, 28, 192)	6144
block_4_depthwise_BN (BatchNormalization)	(None, 28, 28, 192)	768
block_4_depthwise_relu (ReLU)	(None, 28, 28, 192)	0
block_4_project (Conv2D)	(None, 28, 28, 32)	1728
block_4_project_BN (BatchNormalization)	(None, 28, 28, 32)	768
block_4_add (Add)	(None, 28, 28, 32)	0
block_5_expand (Conv2D)	(None, 28, 28, 192)	6144
block_5_expand_BN (BatchNormalization)	(None, 28, 28, 192)	128
block_5_expand_relu (ReLU)	(None, 28, 28, 192)	0
block_5_depthwise (DepthwiseConvolution)	(None, 28, 28, 192)	6144
block_5_depthwise_BN (BatchNormalization)	(None, 28, 28, 192)	128

Layer(type)	Output Shape	Param #
block_5_depthwise_relu (ReLU)	(None, 28, 28, 192)	0
block_5_project (Conv2D)	(None, 28, 28, 32)	6144
block_5_project_BN (BatchNormalization)	(None, 28, 28, 32)	768
block_5_add (Add)	(None, 28, 28, 32)	0
block_6_expand (Conv2D)	(None, 28, 28, 192)	0
block_6_expand_BN (BatchNormalization)	(None, 28, 28, 192)	1728
block_6_expand_relu (ReLU)	(None, 28, 28, 192)	768
block_6_pad (ZeroPadding2D)	(None, 29, 29, 192)	0
block_6_depthwise (DepthwiseConvolution)	(None, 14, 14, 192)	12288
block_6_depthwise_BN (BatchNormalization)	(None, 14, 14, 192)	256
block_6_depthwise_relu (ReLU)	(None, 14, 14, 192)	24576
block_6_project (Conv2D)	(None, 14, 14, 64)	1536
block_6_project_BN (BatchNormalization)	(None, 14, 14, 64)	0
block_7_expand (Conv2D)	(None, 14, 14, 384)	24576
block_7_expand_BN (BatchNormalization)	(None, 14, 14, 384)	1536
block_7_expand_relu (ReLU)	(None, 14, 14, 384)	0
block_7_depthwise (DepthwiseConvolution)	(None, 14, 14, 384)	3456
block_7_depthwise_BN (BatchNormalization)	(None, 14, 14, 384)	1536
block_7_depthwise_relu (ReLU)	(None, 14, 14, 384)	0
block_7_project (Conv2D)	(None, 14, 14, 64)	24576
block_7_project_BN (BatchNormalization)	(None, 14, 14, 64)	256
block_7_add (Add)	(None, 14, 14, 64)	0
block_8_expand (Conv2D)	(None, 14, 14, 384)	24536
block_8_expand_BN (BatchNormalization)	(None, 14, 14, 384)	1536
block_8_expand_relu (ReLU)	(None, 14, 14, 384)	0
block_8_depthwise (DepthwiseConvolution)	(None, 14, 14, 384)	3456
block_8_depthwise_BN (BatchNormalization)	(None, 14, 14, 384)	1536
block_8_depthwise_relu (ReLU)	(None, 14, 14, 384)	0
block_8_project (Conv2D)	(None, 14, 14, 64)	24576
block_8_project_BN (BatchNormalization)	(None, 14, 14, 64)	256
block_8_add (Add)	(None, 14, 14, 64)	0
block_9_expand (Conv2D)	(None, 14, 14, 384)	24576
block_9_expand_BN (BatchNormalization)	(None, 14, 14, 384)	1536
block_9_expand_relu (ReLU)	(None, 14, 14, 384)	0
block_9_depthwise (DepthwiseConvolution)	(None, 14, 14, 384)	3456
block_9_depthwise_BN (BatchNormalization)	(None, 14, 14, 384)	1536
block_9_depthwise_relu (ReLU)	(None, 14, 14, 384)	0
block_9_project (Conv2D)	(None, 14, 14, 64)	24576
block_9_project_BN (BatchNormalization)	(None, 14, 14, 64)	256
block_9_add (Add)	(None, 14, 14, 64)	0
block_10_expand (Conv2D)	(None, 14, 14, 384)	24576
block_10_expand_BN (BatchNormalization)	(None, 14, 14, 384)	1536
block_10_expand_relu (ReLU)	(None, 14, 14, 384)	0
block_5_expand_BN (BatchNormalization)	(None, 28, 28, 192)	128
block_5_expand_relu (ReLU)	(None, 28, 28, 192)	0
block_5_depthwise (DepthwiseConvolution)	(None, 28, 28, 192)	6144
block_5_depthwise_BN (BatchNormalization)	(None, 28, 28, 192)	128
block_5_depthwise_relu (ReLU)	(None, 28, 28, 192)	0
block_5_project (Conv2D)	(None, 28, 28, 32)	6144
block_5_project_BN (BatchNormalization)	(None, 28, 28, 32)	768
block_5_add (Add)	(None, 28, 28, 32)	0

Layer(type)	Output Shape	Param #
block_6_expand (Conv2D)	(None, 28, 28, 192)	0
block_6_expand_BN (BatchNormalization)	(None, 28, 28, 192)	1728
block_6_expand_relu (ReLU)	(None, 28, 28, 192)	768
block_6_pad (ZeroPadding2D)	(None, 29, 29, 192)	0
block_6_depthwise (DepthwiseConvolution)	(None, 14, 14, 192)	12288
block_6_depthwise_BN (BatchNormalization)	(None, 14, 14, 192)	256
block_6_depthwise_relu (ReLU)	(None, 14, 14, 192)	24576
block_6_project (Conv2D)	(None, 14, 14, 64)	1536
block_6_project_BN (BatchNormalization)	(None, 14, 14, 64)	0
block_7_expand (Conv2D)	(None, 14, 14, 384)	24576
block_7_expand_BN (BatchNormalization)	(None, 14, 14, 384)	1536
block_7_expand_relu (ReLU)	(None, 14, 14, 384)	0
block_7_depthwise (DepthwiseConvolution)	(None, 14, 14, 384)	3456
block_7_depthwise_BN (BatchNormalization)	(None, 14, 14, 384)	1536
block_7_depthwise_relu (ReLU)	(None, 14, 14, 384)	0
block_7_project (Conv2D)	(None, 14, 14, 64)	24576
block_7_project_BN (BatchNormalization)	(None, 14, 14, 64)	256
block_7_add (Add)	(None, 14, 14, 64)	0
block_8_expand (Conv2D)	(None, 14, 14, 384)	24536
block_8_expand_BN (BatchNormalization)	(None, 14, 14, 384)	1536
block_8_expand_relu (ReLU)	(None, 14, 14, 384)	0
block_8_depthwise (DepthwiseConvolution)	(None, 14, 14, 384)	3456
block_8_depthwise_BN (BatchNormalization)	(None, 14, 14, 384)	1536
block_8_depthwise_relu (ReLU)	(None, 14, 14, 384)	0
block_8_project (Conv2D)	(None, 14, 14, 64)	24576
block_8_project_BN (BatchNormalization)	(None, 14, 14, 64)	256
block_8_add (Add)	(None, 14, 14, 64)	0
block_9_expand (Conv2D)	(None, 14, 14, 384)	24576
block_9_expand_BN (BatchNormalization)	(None, 14, 14, 384)	1536
block_9_expand_relu (ReLU)	(None, 14, 14, 384)	0
block_9_depthwise (DepthwiseConvolution)	(None, 14, 14, 384)	3456
block_9_depthwise_BN (BatchNormalization)	(None, 14, 14, 384)	1536
block_9_depthwise_relu (ReLU)	(None, 14, 14, 384)	0
block_9_project (Conv2D)	(None, 14, 14, 64)	24576
block_9_project_BN (BatchNormalization)	(None, 14, 14, 64)	256
block_9_add (Add)	(None, 14, 14, 64)	0
block_10_expand (Conv2D)	(None, 14, 14, 384)	24576
block_10_expand_BN (BatchNormalization)	(None, 14, 14, 384)	1536
block_10_expand_relu (ReLU)	(None, 14, 14, 384)	0
block_15_project (Conv2D)	(None, 7, 7, 160)	153600
block_15_project_BN (BatchNormalization)	(None, 7, 7, 160)	640
block_15_add (Add)	(None, 7, 7, 160)	0
block_16_expand (Conv2D)	(None, 7, 7, 960)	153600
block_16_expand_BN (BatchNormalization)	(None, 7, 7, 960)	3840
block_16_expand_relu (ReLU)	(None, 7, 7, 960)	0
block_16_depthwise (DepthwiseConvolution)	(None, 7, 7, 960)	8640
block_16_depthwise_BN (BatchNormalization)	(None, 7, 7, 960)	3840
block_16_depthwise_relu (ReLU)	(None, 7, 7, 960)	0
block_16_project (Conv2D)	(None, 7, 7, 320)	307200
block_16_project_BN(BatchNormalization)	(None, 7, 7, 320)	1280
Conv_1 (Conv2D)	(None, 7, 7, 1280)	409600

Layer(type)	Output Shape	Param #
Conv_1_bn (BatchNormalization)	(None, 7, 7, 1280)	5120
out_relu (ReLU)	(None, 7, 7, 1280)	0
average_pooling2d_2 (AveragePooling)	(None, 1, 1, 1280)	0
flatten (Flatten)	(None, 1280)	0
dense_4 (Dense)	(None, 128)	163968
dropout_2 (Dropout)	(None, 128)	0
dense_5 (Dense)	(None, 2)	258

Total parameters: 2,422,210
Trainable parameters: 164,226
Non-trainable parameters: 2,257,984

References

[1] V. C.-Ch. Cheng *et al.*, "The role of community-wide wearing of face mask for control of coronavirus disease 2019 (COVID-19) epidemic due to SARS-CoV-2," *Journal of Infection*, vol. 81, no. 1, pp. 107–114, 2020, doi: 10.1016/j.jinf.2020.04.024.

[2] A. N. Desai and D. M. Aronoff, "Masks and Coronavirus Disease 2019 (COVID-19)," *JAMA*, vol. 323, no. 20, p. 2103, May 2020, doi: 10.1001/jama.2020.6437.

[3] C. Kenyon, "The prominence of asymptomatic superspreaders in transmission mean universal face masking should be part of COVID-19 de-escalation strategies," *International Journal of Infectious Diseases*, vol. 97, pp. 21–22, 2020.

[4] N. H. L. Leung *et al.*, "Respiratory virus shedding in exhaled breath and efficacy of face masks," *Nature Medicine*, vol. 26, no. 5, pp. 676–680, 2020, doi: 10.1038/s41591-020-0843-2.

[5] Y. Liu, A. A. Gayle, A. Wilder-Smith, and J. Rocklöv, "The reproductive number of COVID-19 is higher compared to SARS coronavirus," *Journal of Travel Medicine*, vol. 27, no. 2, Feb. 2020, doi: 10.1093/jtm/taaa021.

[6] M. P. Fang, Yaqing, Yiting Nie, "Transmission dynamics of the COVID-19 outbreak and effectiveness of government interventions: A data-driven analysis," *Journal of medical virology*, vol. 92, no. 6, pp. 645–659, 2020, doi: 10.1002/jmv.25750.

[7] E. S. K. Lim, Soo, Ho Il Yoon, Kyoung-Ho Song and H. Bin Kim, "Face Masks and Containment of Coronavirus Disease 2019 (COVID-19): Experience from South Korea," *The Journal of Hospital Infection*, 2020, doi: 10.1016/j.jhin.2020.06.017.

[8] S. Esposito, N. Principi, C. C. Leung, and G. B. Migliori, "Universal use of face masks for success against COVID-19: evidence and implications for prevention policies," *European Respiratory Journal*, p. 2001260, Jan. 2020, doi: 10.1183/13993003.01260-2020.

[9] 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*, vol. 8, no. 12S, pp. 2278–3075, 2019.

[10] H. Panwar, P. K. Gupta, M. K. Siddiqui, R. Morales-Menendez, and V. Singh, "Application of deep learning for fast detection of COVID-19 in X-Rays using nCOVnet," *Chaos, Solitons & Fractals*, vol. 138, p. 109944, 2020, doi: 10.1016/j.chaos.2020.109944.

[11] F. Sultana, A. Sufian, and P. Dutta, "Advancements in Image Classification using Convolutional Neural Network," in *2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, 2018, pp. 122–129, doi: 10.1109/ICRCICN.2018.8718718.

[12] H. Shin *et al.*, "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016, doi: 10.1109/TMI.2016.2528162

[13] M. Coşkun, A. Uçar, Ö. Yildirim, and Y. Demir, "Face recognition based on convolutional neural network," in *2017 International Conference on Modern Electrical and Energy Systems (MEES)*, 2017, pp. 376–379, doi: 10.1109/MEES.2017.8248937.

[14] Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[15] Wang *et al.*, "Residual attention network for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3156–3164.

[16] Z. Zhao, P. Zheng, S. Xu, and X. Wu, "Object Detection With Deep Learning: A Review," *IEEE Transaction on neural networks and learning systems*, vol. 30, no. 11, pp. 3212–3232, 2019, doi: 10.1109/TNNLS.2018.2876865.

[17] Redmon, Joseph, S. Divvala, R. Girshik, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[18] A. Kumar, A. Kaur, and M. Kumar, "Face detection techniques: a review," *Artificial Intelligence Review, Springer*, vol. 52, no. 2, pp. 927–948, 2019, doi: 10.1007/s10462-018-9650-2.

[19] A. S. M. and V. M. M. Sivaram, V. Porkodi, "Detection of Accurate Facial Detection using Hybrid Deep Convolutional Recurrent Neural Network," *ICTACT Journal of Soft Computing*, vol. 9, no. 2, 2019, doi: 10.21917/ijsc.2019.0256.

[20] M. Khan, S. Chakraborty, R. Astya, and S. Khepra, "Face Detection and Recognition Using OpenCV," in *2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, 2019, pp. 116–119, doi: 10.1109/ICCCIS48478.2019.8974493.

[21] T. Meenpal, A. Balakrishnan, and A. Verma, "Facial Mask Detection using Semantic Segmentation," in *2019 4th International Conference on Computing, Communications and Security (ICCCS)*, 2019, pp. 1–5, doi: 10.1109/CCCS.2019.8888092.

[22] M. Jiang, X. Fan, and H. Yan, "RetinaMask: A Face Mask detector." 2020, [Online]. Available: http://arxiv.org/abs/2005.03950.

[23] T. Zhang, J. Li, W. Jia, J. Sun, and H. Yang, "Fast and robust occluded face detection in ATM surveillance," *Pattern Recognition Letters*, vol. 107, pp. 33–40, 2018, doi: 10.1016/j.patrec.2017.09.011.

[24] H. Baojin, "Real-World Masked Face Dataset, RMFD." Real-World Masked Face Dataset, RMFD.

[25] "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017, [Online]. Available: https://arxiv.org/abs/1704.04861.

[26] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.

[27] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[28] "How to train a Face Detector," https://github.com/opencv/opencv/blob/4.0.0-beta/samples/dnn/face_detector/how_to_train_face_detector.txt.

[29] "deploy.prototxt," https://github.com/opencv/opencv/blob/4.0.0-beta/samples/dnn/face_detector/deploy.prototxt.

[30] Y. Jia *et al.*, "Caffe: Convolutional Architecture for Fast Feature Embedding," in *Proceedings of the 22nd ACM International Conference on Multimedia*, 2014, pp. 675–678, doi: 10.1145/2647868.2654889.

[31] D. L. Streiner and J. Cairney, "What's under the ROC? An Introduction to Receiver Operating Characteristics Curves," *The Canadian Journal of Psychiatry*, vol. 52, no. 2, pp. 121–128, Feb. 2007, doi: 10.1177/070674370705200210.

[32] Yeh and Shi-Tao, "Using trapezoidal rule for the area under a curve calculation," *Proceedings of the 27th Annual SAS User Group International* 2002.

Anubha Bhaik

Anubha Bhaik has recently completed her B.Tech in Computer Science & Engineering from Jaypee University of Information Technology, India. She will be pursuing her Master's in Computer Science at the University of Florida, USA. Her interests lie in deep learning, computer vision and data science. She is an experienced team lead with a demonstrated record of contributing to the research industry.

Vaishnavi Singh

Vaishnavi Singh has recently completed her B.Tech in Computer Science & Engineering from Jaypee University of Information Technology, India. She enjoys figuring out the different building blocks of the technical world and rearranging them to discover new possibilities. She is passinate about exploring and applying various fields of computer science to develop adept solutions for the real-world problems. Her interests lie in deep learning, computer vision, software engineering and database systems. She has a significant history of contributions to the research industry, whereas majority of her work is related to solving the challenges of COVID-19.

Ekta Gandotra

Ekta Gandotra is currently working as Assistant Professor in the Department of Computer Science & Engineering at Jaypee University of Information Technology, Waknaghat (India). She has completed her Ph.D. in Computer Science and Engineering from PEC University of Technology, Chandigarh (India). She has around 12 years of teaching and research experience. Her research areas include network & cybersecurity, malware threat profiling, cyber threat intelligence, machine/deep learning, and big data analytics.

Deepak Gupta

Deepak Gupta is working as Assistant Professor in the Department of Computer Science & Engineering at Jaypee University of Information Technology, Waknaghat (India). He has completed his Ph.D. in Computer Science & Engineering from Thapar Institute of Engineering and Technology (Deemed to be University), Patiala (India). Prior to his foray into academia, he worked in IT industry for a decade performing different roles in software product development and program management. In all, he has more than 20 years of rich experience in IT industry and academics. His research interests include big data analytics, machine/deep learning, cybersecurity, and programming languages.

Adaptive Deep Learning Detection Model for Multi-Foggy Images

Zainab Hussein Arif[1,2], Moamin A. Mahmoud[1], Karrar Hameed Abdulkareem[3]*, Seifedine Kadry[4,5,6], Mazin Abed Mohammed[7,8], Mohammed Nasser Al-Mhiqani[9], Alaa S. Al-Waisy[2], Jan Nedoma[8]

[1] College of Computing and Informatics, Universiti Tenaga Nasional, 43000 Kajang, Selangor (Malaysia)
[2] Computer Technologies Engineering Department, Information Technology Collage, Imam Ja'afar Al-Sadiq University, Baghdad (Iraq)
[3] College of Agriculture, Al-Muthanna University, Samawah 66001 (Iraq)
[4] Department of Applied Data Science, Norrof University College, 4608 Kristiansand (Norway)
[5] Artificial Intelligence Research Center (AIRC), College of Engineering and Information Technology, Ajman University, Ajman, (United Arab Emirates)
[6] Department of Electrical and Computer Engineering, Lebanese American University, Byblos (Lebanon)
[7] College of Computer Science and Information Technology, University of Anbar, 11, Ramadi, Anbar (Iraq)
[8] Department of Telecommunications, VSB-Technical University of Ostrava, 70800 Ostrava (Czech Republic)
[9] School of Computer Science and Mathematics, Faculty of Natural Sciences Keele University (KU), Keele (United Kingdom)

Received 25 February 2021 | Accepted 3 February 2022 | Published 30 November 2022

Abstract

The fog has different features and effects within every single environment. Detection whether there is fog in the image is considered a challenge and giving the type of fog has a substantial enlightening effect on image defogging. Foggy scenes have different types such as scenes based on fog density level and scenes based on fog type. Machine learning techniques have a significant contribution to the detection of foggy scenes. However, most of the existing detection models are based on traditional machine learning models, and only a few studies have adopted deep learning models. Furthermore, most of the existing machines learning detection models are based on fog density-level scenes. However, to the best of our knowledge, there is no such detection model based on multi-fog type scenes have presented yet. Therefore, the main goal of our study is to propose an adaptive deep learning model for the detection of multi-fog types of images. Moreover, due to the lack of a publicly available dataset for inhomogeneous, homogenous, dark, and sky foggy scenes, a dataset for multi-fog scenes is presented in this study (https://github.com/Karrar-H-Abdulkareem/Multi-Fog-Dataset). Experiments were conducted in three stages. First, the data collection phase is based on eight resources to obtain the multi-fog scene dataset. Second, a classification experiment is conducted based on the ResNet-50 deep learning model to obtain detection results. Third, evaluation phase where the performance of the ResNet-50 detection model has been compared against three different models. Experimental results show that the proposed model has presented a stable classification performance for different foggy images with a 96% score for each of Classification Accuracy Rate (CAR), Recall, Precision, F1-Score which has specific theoretical and practical significance. Our proposed model is suitable as a pre-processing step and might be considered in different real-time applications.

Keywords

Deep Learning, Fog Detection, Foggy Image, Multi-Fog, Multi-Class Classification.

DOI: 10.9781/ijimai.2022.11.008

I. Introduction

Fog recognition and detection of objects in foggy weather condition is important for many applications such as transportation systems, and autonomous driving [1]-[3]. Foggy conditions can cause a serious traffic safety problem if the autonomous car on-board smart sensors fail to detect other cars or pedestrians [4]-]5]. Therefore, development of artificial intelligence methods and smart sensing technologies for fog recognition is important in machine vision.

As an atmospheric effect, fog creates a grey color over the scene, thereby degrading visibility in outdoor scene images [6]. Besides, the fog has been regarded as one of the main sources of accidents that occur in different environments like air, underwater, and over-land [7]. Fog is formed when the light which propagates through the atmosphere is scattered by particles like moisture and smoke; these particles are normally dispersed by air [8]. To eliminate the effects of the fog inside the obtained image, the art of dehazing is employed. The process of dehazing involves the elimination of the effects of fog

* Corresponding author.
E-mail address: khak9784@mu.edu.iq

from outdoor images, as well as the restoration of fidelity details. Conceptually, the technique of dehazing image which is also referred to as "defogging" or "fog removal", is an image enhancement that involves the elimination of undesirable visible effects from an image [9]. However, this technique is not the same as the conventional noise elimination technique and contrast enhancement techniques, since the degradation to image pixels that is induced by the presence of fog is dependent on the distance between the object and the acquisition device and the regional density of the fog [10]. The wide range of colors is overshadowed by the effect which the fog has on image pixels [11]. The presence of particles in the atmosphere, makes the execution of computer vision tasks in the presence of fog difficult and effective, thereby leading to the production of heavily degraded images [12]-[14]. Thus, it becomes crucial to subject the degraded images to process of defogging [15]-[17].

According to the popularity of the traditional machine learning techniques in vast application areas, such techniques have applied widely in the image defogging domain to tackle issues related to image denoising [18], image quality assessment [19], image segmentation [20], and image classification or detection [21]. Comparing with traditional techniques, deep learning-based methods have made remarkable progress in image dehazing problems, especially in image denoising [22] and image classification [23]. Most recently, several deep learning-based approaches have proposed that demonstrate more robustness as compared to the traditional non-learning-based methods [24]. With the presence of a deep learning model [25], complex networks can be created so that the problem of image classification can be solved. Such a problem is often solved using Convolutional neural networks (CNNs), where complex networks serve as a collection of feature extractors that are usually somewhat generic, and relatively free of any classification task [26]. Naturally, the performances of these algorithms are superior to those of conventional manually crafted priors-based techniques by a significantly large margin. This can be attributed to the fact that informative and useful features can be extracted by CNNs from large amounts of images with strong generalization capability [22]. More so, due to the automatic process of extracting features, human supervision is not required [23].

Determining the presence and magnitude of fog has a great enlightening effect on the process of image defogging. Since it is possible to perform subsequent processing selectively, images can be better understood [27]. The process of detecting fog in image can be thought as a classification problem [21], [28]. However, in this study we do not discriminate between classification and detection terms. Fog has a significant impact on the image and cause different types of noise such as contrast, color, and structure distortion [29]. Usually, the classification of foggy images is based on machine learning techniques. The machine learning models need to predict the image type according to relevant features with corresponding environment characteristics. Furthermore, foggy image features could be varied according to density of fog or foggy scene type. For example, in the inhomogeneous foggy scene, the distribution of the fog is not equal in whole image, while in the homogenous foggy scene the fog level almost the same in whole image. Thus, a significant challenge can be seen in terms of classification of image according to more complex foggy image features. Furthermore, a comprehensive platform that covers all characteristics of foggy images in the classification of foggy images should be developed.

Irrespective of the absence or presence of fog, the direct defogging algorithms are applied to images or videos. However, when real-world applications are involved, it is essential to determine if there is a need for the processing of the obtained image in the given environment by a defogging algorithm. This is because, without any judgment, the use of a defogging algorithm to restore an image may worsen the visibility of the image as compared to the original image. Besides, a particular noise could be added to a picture that has no fog when the defogging algorithm is used in processing degraded images. This, in turn, distorts the image to a certain level. Other demerits of using a defogging algorithm include low efficiency of processing, unconducive for image recognition, and consumes much time. Thus, it is crucial to effectively determine the presence of fog in an image before it is subjected to the process of defogging. This implies that the classification and recognition of the acquired image before processing is practically significant and valuable for application; this must be considered before improving the efficiency of image processing [23],[29],[30].

Additionally, the extant technical theory is restricted to fog image processing, and as such, it is difficult to make a valid classification of the image as non-foggy or foggy. As a result, relying solely on subjective judgement will not be able to meet the demands of real-time and batch processing. Consequently, having a method for accurately determining the state of the image is critical [31]. Also, another problem associated with fog is the detection of context information that describes the condition of the environment such as snow, fog, or rain. This problem is well established in the extant literature [30].

This paper analyzes the suitability of four deep learning models (Xception, VGG-16, Inception-V3, ResNet-50) for fog detection. The methodology is evaluated based on the dataset of collected foggy images. The contribution of this study can be seen as follow:

- Based on seven public datasets, the first dataset that includes four foggy scenes (inhomogeneous, homogeneous, dark, and sky foggy scene) is presented to the research community and it is openly available from https://github.com/Karrar-H-Abdulkareem/Multi-Fog-Dataset.
- The first adaptive multi-class study is presented based on the classification of four foggy scenes which are inhomogeneous, homogeneous, dark, and sky foggy scenes. The adaptive term refers to capability of classification model to classify different foggy scenes.
- We adopted the ResNet-50 model for the classification of multi-class foggy images.

The remaining parts of the paper present an overview of the related works (Section II), present the methodology (Section III), present and discusses the results (Section IV), discuss the application scenarios and the limitations of our approach (Section V), and state the conclusions (Section VI).

II. Related Works

In general, the foggy scene is divided into two types, a scene based on fog level density and scenes based on fog type where different characteristics are recognized such as time, light source, and so on. Scenes based on fog type have different categories. First, an inhomogeneous foggy scene which is referred to an image that usually contains an uneven amount of fog that is distributed in the entire image [32]. Second, a homogenous foggy scene is assumed that amount of fog is equally distributed in the entire image [7]. Third, the dark foggy scene is common, including visible lights (artificial lights) sources with varying colors besides the presence of fog. The light sources also often introduce noticeable amounts of glow that are not present in daytime fog [33]. Fourth, sky foggy scenes where the color of the sky is usually very similar to the atmospheric light in a fog image [34], so the pixel with the highest intensity might correspond to a bright object rather than to the airlight [35].

Most of the current classification models in the image defogging area are based on two insights. First, classification models are based on binary classification, where the main target of these models is to

distinguish between two types of images, i.e., foggy and non-foggy, respectively [28]. Second, classification is based on multi-class models which is the main purpose of these models to classify images based on fog density [21]. Many studies have addressed the problem of foggy image classification, as illustrated in Table I.

Many techniques have been used to tackle the problem of foggy image classification. Most classification techniques belong to traditional machine learning approaches rather than deep learning. Even with new trends of transfer learning, many investigated studies prefer to use hand-crafted features, especially in terms of color and gradient features. Furthermore, the process of feature extraction quite complex in some studies. The SVM technique is widely used as a classification model to characterize foggy images since the high accuracy can be obtained [27]. All binary classification studies focused on the general foggy images classified as foggy and non-foggy images. Furthermore, the binary classification has generalized to the more specific foggy scene such as non-sky and foggy sky images. On the other hand, multi-class classification also has been used to classify the images based on fog levels such as light, medium, and heavy foggy scenes. Only one study [28] has combined each of binary and multi-class classification. However, many studies have success to classify the daytime while minimum proposed algorithms have ability to work in the night-time foggy images.

Kaiming et al. [39] have presented a residual learning framework where the layers learn residual functions with respect to the inputs received instead of learning unreferenced functions. This model allows the training of profound networks up to more than 1000 layers [39]. It is a well-established fact that the performance of a network is determined by its depth. In the area of computer vision, use of deeper networks is employed. Nevertheless, the training of a deeper network is difficult because of the problem of gradient vanishing, which is difficult to combat. Therefore, for this problem to be addressed, the use of ResNet can be employed [39] because it offers a training framework that can simplify the training of networks that are reasonably deeper than the previously used ones [40]. The motivation for this is the findings of scientific experiments that have shown that the level of training error increases as more layers are added. Conceptually, an increase in the number of layers should lead to increased modeling efficiency of Neural Networks, thereby preventing the occurrence of higher training error. This is attributed to the fact that after the propagation of gradients in many layers, they (gradients) vanish. Rather than enabling the uninterrupted flow of a portion of earlier information to later layers via Highway Networks through the addition of parameterized gating functions [41], the authors in [40] suggested that shortcut connections with identity should be simply added to the networks. Moreover, ResNet-50 has never been used before for the classification of foggy images, furthermore, according to the mentioned advantages, ResNet-50 has been employed for the classification target of four foggy image classes in this study.

Finally, the findings of the examined research have shown three major difficulties. First, most of the proposed models have lack of evaluation based on the benchmarked dataset and prefer to use the private dataset that is based on the authors' own experiment, which raises a question about the possibility of existed studies if can be generalized for more challenging scenarios in real-time applications. Furthermore, to the best of our knowledge, there is no such available public dataset that has been covered each inhomogeneous, homogeneous, dark, and sky foggy scene. Second, all mentioned studies have ignored the multi-class classification of foggy images based on the different scene types such as the inhomogeneous, homogeneous, dark, and sky foggy scene. Third, according to [23] the authors have shown the advantages of using deep learning models, especially in the matter of accuracy comparing with traditional machine learning models in the binary classification of foggy images. However, to our knowledge, no study has applied a deep learning model for classification each of inhomogeneous, homogeneous, dark, and sky foggy image as multi-class classification task in the image defogging domain, which we consider to be a theoretical gap.

III. Methodology

This section presents the full details for the proposed model into different phases as follows:

A. Data Collection Phase

As mentioned in the previous section, the lack of a public dataset based on four foggy scenes as well as all classification studies have been conducted based on private datasets which raise a significant challenge in this study. The main aim of this phase is to collect foggy images belong to four foggy scenes; namely, homogenous, inhomogeneous, dark, and sky foggy scene (see Fig.1). The baseline for collected images is the definition for each foggy scene that is already mentioned in section two, where each image scene has different characteristics. Furthermore, the source of collected images is based on seven datasets that can be identified as follows:

- Dehazing using color-lines dataset [42]: This dataset contains eleven foggy images; these images belong to indoor and outdoor foggy scenes. Furthermore, all images with PNG format, but with different dimensions, for instance, 1024, 1200, and so on.
- I-Haze dataset [43]: a dataset, in which 35 pairs of foggy and corresponding haze-free (ground-truth) indoor images are contained therein. Unlike the majority of the available dehazing databases, the real haze has been used in generating foggy images; the use of a professional haze machine was employed in generating the images. A MacBeth color checker was included in all the scenes so that color calibration can be eased while the evaluation of the dehazing algorithm can be improved. Besides, since the capturing of images is done within a controlled environment, both foggy and haze-free images are captured under the same lighting conditions. All images obtained have the JPG format with 4675 × 2833 px dimensions.
- Kede dataset [44]: this dataset contains a total number of 225 images, out of which 200 are defogged, and 25 are foggy. The 225 images were divided into 9 image sets with 25 images each. These images are provided to cover diverse outdoor scenes and different

(a) Inhomogeneous (b) Homogeneous

(c) Dark (d) Sky

Fig. 1. Foggy scenes are based on different fog types.

TABLE I. Existing Studies on Classification for Foggy Images.

Ref	Model	Features	Advantage	Classification type	Disadvantage	Scene type
[31]	SVM (linear kernel)	Hand-crafted features	New indicators to distinguish between foggy and non-foggy images	Binary	No comparative scenario provided	Foggy and non-foggy
[36]	SVM (linear kernel)	Hand-crafted features Scattering Model	Simple implementation	Multi-class	Features aren't good enough to describe the whole information of images	Clear, light foggy, medium foggy, and heavy foggy
[30]	SVM	Hand-crafted features	Sample size has a very small effect on speed	Binary	Lack of comprehensive evaluation scenario	Foggy and non-foggy
[21]	SVM (RBF)	Hand-crafted features	Hight potential for fog detection on daytime images	Multi-class	Doubtable to efficiently works in dark foggy scene	Excluded, No Fog, Low Fog, Fog and Dense Fog
[22]	AlexNet	Deep features	Process of feature extraction is automatic and needs no supervision	Binary	Number of hyper tuning parameters quite high	Foggy and non-foggy
[28]	SVM (LDA)	Hand-crafted features	Works well for day-time scenes	Binary	Limited only to grayscale images	Foggy and clear images in daytime
[28]	SVM (LDA)	Hand-crafted features	Works well in night-time scenes	Binary	Limited only to grayscale images	foggy and clear images in night-time
[23]	Deep learning	Hand-crafted features	Process of feature extraction is automatic and needs no supervision	Binary	Number of hyper tuning parameters quite high	Clear and foggy image
[27]	SVM (RBF)	hand-crafted features	Efficient for fog density classification	Multi-class	Complex features extraction	Fog-free, thin fog and dense fog image
[36]	19 classification techniques	Hand-crafted features	Comprehensive evaluation	Binary	Too many features extracted	Foggy and non-foggy
[37]	SVM	Hand-crafted features	Efficient to work for large sky foggy classification	Binary	Lack of evaluation scenario	Sky and non-sky foggy image
[38]	Two SVM (RBF)classifiers	Hand-crafted features	Better performance both on the detection rate and the misclassification	Binary	Performing poorly for night scenes	Sky and non-sky foggy image

degrees of haze thickness. These include humans, animals, plants, architecture, landscapes, statics, traffics, and night scenes. Many images were captured in the real world, but the simulation of the three hazy and static objects was done uniformly. All the images are provided in JPG with 4675 x 2833 dimensions. The evaluation of the image involved the participation of 24 naïve observers, out of which 12 were female and 12 males between the age of 22 and 28.

- LIVE Image Defogging Database: The LIVE Image Defogging Database proposed by [45], this dataset has been used widely in many evaluation scenarios in image dehazing domain such as [7], [11],[46]. A total of 1100 natural fog-free, foggy, and test images have been presented by this database. Here 100 colored images were selected to provide adequately diverse images, and fog density from newly recorded foggy images, well-established foggy test images, and corresponding defogged images. The images were of different sizes ranging from 425 × 274 to 1024 × 768 pixels. Besides, few foggy scenes like dark foggy scenes, homogenous scenes, and inhomogeneous foggy scenes were contained in the images. However, all provided images are in different formats and dimensions.
- O-haze dataset [44]: O-haze is a dataset composed of 45 kinds of outdoor scenes. It contains pairs of corresponding haze-free images and real foggy images. Practically, the capturing of the foggy images was done in real haze using professional haze machines. All the images in this dataset represent the same visual content recorded under foggy and haze-free conditions and the same lighting condition. The significance of O-HAZE has been proven by using it to make a comparison of a representative set of state-of-the-art dehazing methods. The comparison involved the use of conventional image quality metrics like SSIM, PSNR, and CIEDE2000. Through this comparison, the shortcomings of the current techniques are uncovered, and some based on the uncovered shortcomings, some of their underpinning hypotheses are questioned. Nevertheless, all the images have been provided in varying dimensions and formats.
- RESIDE Database: this dataset has been proposed by [47]; it has also been widely used in many evaluation studies [48], [49] and [50]; it features a large-scale synthetic training set and two different sets of designed quality evaluations, respectively. RESIDE has diverse data sources and image contents. This dataset has indoor and outdoor images, clear and foggy images, with more than 12,000 real-world images. However, all provided images are in different formats and dimensions.
- Foggy Image dataset (FI): this dataset was proposed by [51] which contained 200 foggy images. The corresponding manual labeled ground truths have been provided in this dataset. This dataset was mainly used for object detection in foggy weather. However, all provided images are in the PNG format but with different dimensions.

Furthermore, the data collection process time depends on the volume of images in each dataset. The final set of collected images depend on the availability of four mentioned foggy scenes in each defined dataset. However, the number of collected images may present a negative effect on the classification task, especially with deep learning models, especially when the number of collected images for the training dataset is small. Thus, a recommended solution is needed to avoid such a challenge.

B. Classification of Foggy Image Phase

This phase focuses on all processes relevant to the multi-class classification model such as pre-processing, classification model, and evaluation process (see Fig. 2).

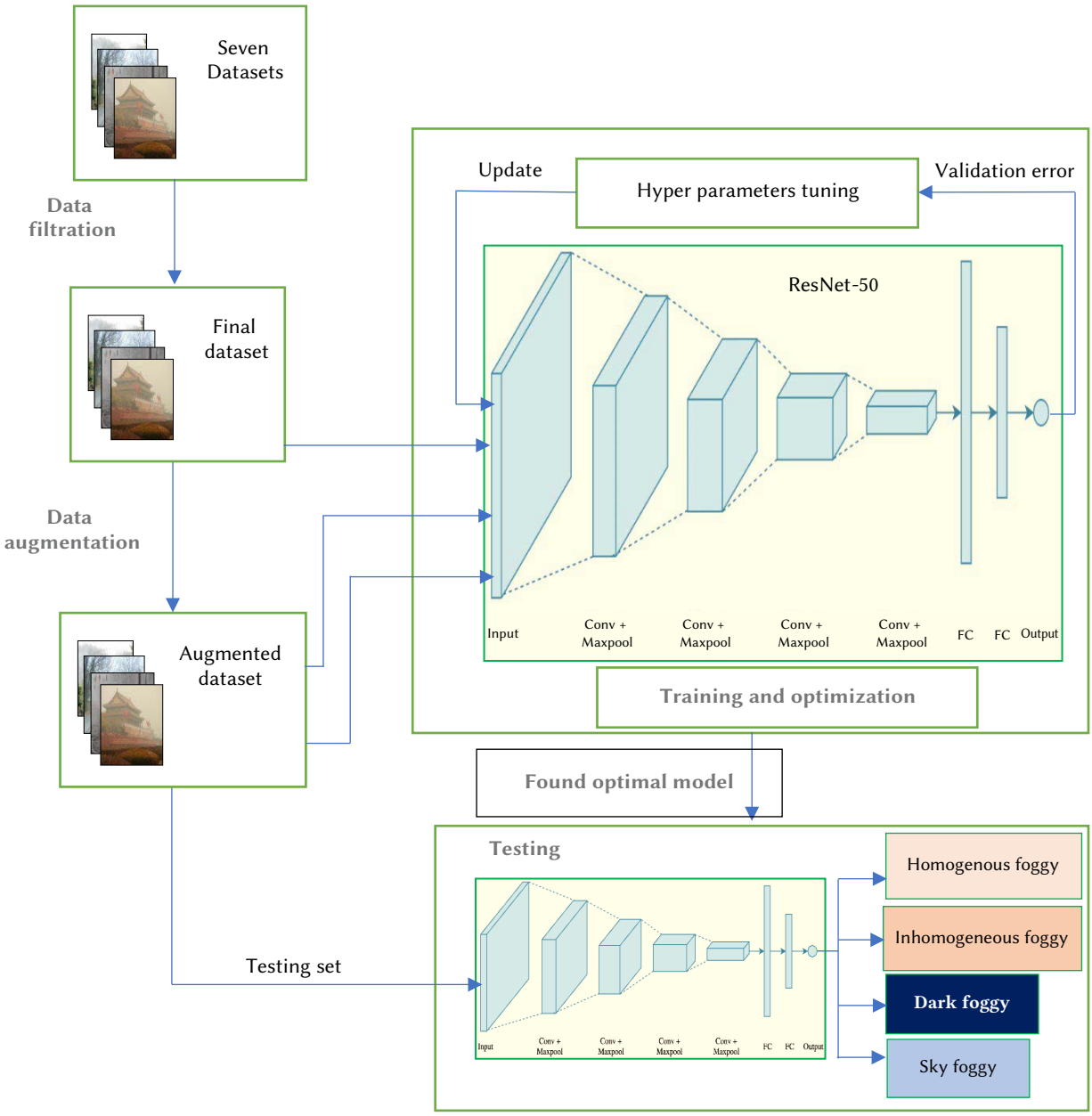


Fig. 2. New classification model for four multi-fog classes.

C. Data Pre-Processing Stage

Before training the model, it is fundamental to apply pre-processing techniques to avoid distorted information. Thus, it allows a correct and more straightforward evaluation of data through the network. In this research, the images of the foggy scenes are collected from different sources. The images are captured by different types of equipment and contain different acquisition parameters. As a result, there exist considerable variations in the intensity of the images. However, the proposed CNN model implements several standard pre-processing procedures to ensure that the generalization of the CNN models is not negatively affected, as follows:

- **Resizing**: we need first to acquire a constant dimension because all images in this dataset vary in dimension and resolution (365×465 to 1125×859 pixels). Subsequently, all the images are scaled to specific pixels based on the corresponding CNN model (e.g., ResNet_50 224×224 pixels).
- **Normalization:** In the normalization part, to set the scaling limit, we use a precalculated mean subtraction of the ImageNet database to normalize the intensity values [52]. Then we scale the intensity values from [0, 255] to the intensity range of [0, 1] using the min-max normalization formula.
- **Image data shuffling:** Random shuffling of data is a standard procedure in all machine learning pipelines, and image classification is not an exception; its purpose is to break possible biases during data preparation - e.g., putting all the sky foggy images first and then the homogenous ones in a foggy image classification dataset.
- Furthermore, make all input data in a similar image format where all foggy images will be set to PNG format.
- **Data augmentation:** A data augmentation procedure is suggested to avoid the overfitting problem during the learning process especially with a small dataset [53,-57] and increase the

generalization ability of the last trained model, we will perform a detailed offline augmentation over the collected foggy images in our dataset, the listed augmentation technics will apply to original images:

1. Rotation with range = 5 degrees.
2. Zoom With range=0.1.
3. Width Shifting with range=0.1.
4. Height Shifting with range=0.1.
5. Horizontal flip.

Zoom is a float value for the zoom_range parameter that takes a lower limit and an upper limit. The shift technique helps in improving those images that are not properly positioned. The values specified are either specified in the form of percentage or integer (in our paper we use percentage). Random flip is a Boolean input that randomly flipping half of the images horizontally or vertically. We have excluded some of the augmentation techniques, for instance, brightness since the most defect in the foggy image is the low contrast, so applying such augmentation technique may affect the real characteristics of the images thus affects all other analysis processes.

D. Image Classification Based on ResNet50 Model

In this paper, ResNet-50 is applied for several reasons, including: (i) to reduce the training time required to obtain the last trained model, and low training error rate especially when the depth of network is increased, (ii) to increase the prediction accuracy of the proposed Multi-class foggy images classification model. To mention, the process of detection fog depends on different factors. First, scene type is the first indicator for detection for instance; image that taken at the night with presence of artificial light considered as dark foggy image. Second, image that contained fog in small portion or overall image also consider as foggy image.

Here, we employed an efficient pre-trained ResNet-50 model. Residual Network (ResNet) one of the most powerful deep CNNs. ResNet is similar to other CNNs, which have convolutional, pooling, activation maps, and fully connected layers stacked sequentially one over the other. The only main difference between ResNet and other CNNs is the identity connection, which is originating from the input layer to the end of the residual block (see Fig. 3-b).

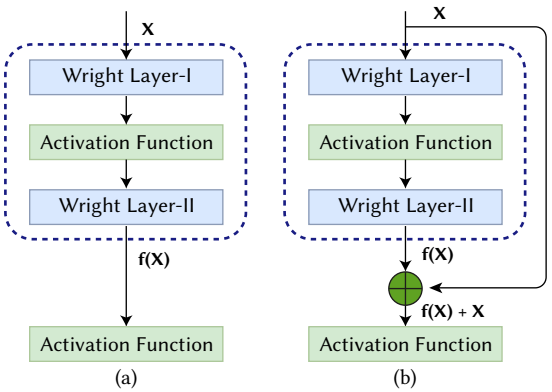


Fig. 3. Concept of ResNet-50 model.

We simply explain the theory of the ResNet-50 below. We explicitly make these layers fit a residual mapping, instead of each few stacked layers directly fit a desired underlying mapping H (x). We let the stacked nonlinear layers fit another mapping of F (x):= H (x) – x. The original mapping is recast into F (x) + x. The formulation of F (x) + x can implement by feedforward neural networks with shortcut connections. The shortcut method uses the connection that skips one or more layers, which simply performs identity mapping. Their outputs are added to the outputs of the stacked layers. Identity shortcut connections do not add any extra parameter or computational complexity. The network can be trained end-to-end by stochastic gradient descent (SGD) with backpropagation.

The main steps of the proposed training methodology can be summarized as follows:

1. Dividing the augmented database into three different sets: Training set, Validation set, and Test set.
2. Select initial values for a set of hyper-parameters (e.g., learning rate, momentum, weight decay, etc.).
3. Training the ResNet-50 using the training set and the hyper-parameters set in step 2.
4. Using the validation set to evaluate the performance of the ResNet-50 during the training process.
5. Repeating steps 3 through 4 for 20 epochs.
6. Selecting the best-trained model with minimal error rate on the validation set.

E. Evaluation Phase

Based on the testing set and the best-trained model, the evaluation process will be conducted. To evaluate the quantitative performance of the proposed model, such as evaluation metrics Classification Accuracy Rate (CAR) , Precision, Recall, and F1 score are computed to validate the efficiency and reliability of the proposed system using the testing set. Furthermore, all mentioned metrics are calculated based on the weighted average value have chosen for the multi-class classification setup, which preferable if you suspect there might be a class imbalance (i.e., you may have many more examples of one class than of other classes). Weighted-avg is calculating metrics for each label, and find their average weighted by support (the number of true instances for each label). Each of weighted CAR, Precision, Recall, and F1 score is calculated according to equations [1]:

$$CAR = \frac{\sum_{i=1}^{m} \frac{tp_i + tn_i}{tp_i + fn_i + fp_i + tn_i}}{m} \quad (1)$$

$$Precision = \frac{\sum_{i=1}^{m} tp_i}{\sum_{i=1}^{m} (tp_i + fp_i)} \quad (2)$$

$$Recall = \frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l} (tp_i + fn_i)} \quad (3)$$

$$F\_score = \frac{\sum_{1}^{m} |y_i| \frac{2tp_i}{2tp_i + fp_i + fn_i}}{\sum_{1}^{m} |y_i|} \quad (4)$$

If m is the total number of classes in the dataset, then i value from 1 to m. Furthermore, to compare the performance of the proposed detection model, the process of evaluation will be conducted within three deep learning models, namely, Xception, VGG-16, Inception-V3.

IV. Results

This section presents in detail the experimental results, analyses, and discussions towards the accomplishments of the proposed multi-class image defogging classification model based on multi-fog types. The remaining of this section consists of five broad sections and is organized as follows in line with the proposed methodology discussed in section 3. Subsection 4.1 discusses the results of data collection. Subsection 4.2 presents the foggy images of classification results.

A. Data Collection

The total images are 1166 used in this study based on the main foggy scene categories are homogeneous scene, inhomogeneous scene,

TABLE II. The Distribution of Foggy Images Based on Eight Data Resources

Foggy Scene	Dehazing using Color-Lines	I-HAZE	IVCDehazing (Kede)	LIVE Image Defogging Database	O-HAZE	RESIDE	Saliency Detection Based (FI)	Internet	Total
Homogeneous scene	11	30	3	114	45	41	12	-	256
Inhomogeneous scene	-		-	37		31	-	105	175
Dark scene	-	-	3	13	-	257	2	-	274
Sky scene	-	-	2	62	-	384	15	-	463
Total	11	30	8	226	45	713	29	105	1166

dark scene, and sky scene. The details of the foggy scene categories are presented in Table II. We labeled the images manually according to the class definition.

Based on Table II, for every dataset, the main processes of the proposed methods are evaluated and analyzed. The foggy scene with a large dataset that is used in the experimental results is the sky scene with 463 images, followed by that dark scene with 274 images. Furthermore, homogeneous scene includes 256 images, and inhomogeneous scene contains 175 images with total cases are 1166 images. The homogeneous scene as having 114 images in LIVE Image Defogging Database from 256 as total, while IVCDehazing (Kede) contains 3 images only. The other distributed in the datasets for the homogeneous scene are Dehazing using Color-Lines 11 images, I-HAZE 30 images, O-HAZE 45 images, RESIDE 41 images, and Saliency Detection Based (FI) 12 samples.

The inhomogeneous scene has 105 images for internet source from 175 as total in this class, while LIVE Image Defogging Database contains 37 images only. The other distributed in the datasets for the inhomogeneous scene is RESIDE 31 images. For Dark scene have 257 images for RESIDE database from 274 as total in this class, while Saliency Detection Based (FI) Database contains 2 images only, LIVE Image Defogging Database contains 13 images, and IVCDehazing (Kede) database contains three samples. Finally, Sky scene has 384 images for RESIDE database from 463 as total in this foggy scene, but IVCDehazing (Kede) database contains 2 images only, LIVE Image Defogging Database involve 62 images, and Saliency Detection Based (FI) contains 15 images as shown in Table II. The most source that has contributed to our collected data is RESIDE dataset, while each of Dehazing using Color-Lines and IVCDehazing datasets have the lowest contribution.

However, the variation of samples or cases is affecting the training and testing process. It is observed that some datasets have good samples in the specific foggy scene, but others do not have enough samples for the training and testing process. Furthermore, the most type of foggy class that is prone to create an imbalanced dataset is the inhomogeneous foggy image class followed by each of the homogenous and dark foggy scene classes. Thus, the dataset with mentioned flaws will present a significant challenge when the real experiment is conducted, that is why an efficient solution is needed to adopt for tackling such issues.

B. Foggy Images of Classification Results

In the domain of artificial intelligence, especially deep learning utilization, big data is the main fundamental to supporting the learning task of the features of objects in image processing and classification.

The network training demands big data to support the feature extraction process to get better features of the objects. Data augmentation and data equilibrium are used to avoid the problem of few image samples that are not enough for the training process. Data augmentation is important to transforming the training samples or images and generating new images by confirmed techniques. Using data augmentation, the original database can be enhanced and increased. Also, notably in the training work, assist in preventing the overfitting issue. Thus, it is important to adopt mentioned processes for the detection and classification performance of the proposed model.

In the collected foggy images, most of these images are extracted from different datasets with different characteristics and numbers. This type of database cannot be applied directly to the training process, especially using deep learning methods. However, the scale of this database classified based on the original images is very small to qualify the sample number of images that can be used for the training process. Thus, the data augmentation must be performed on the databases to qualify the samples used and the need to ensure a useful feature extracted. To provide a more efficient training process with adequate deep learning results and avoid previously mentioned issues with the collected dataset, a data augmentation approach has been adopted in our study. Several data augmentation techniques have been applied, such as Rotation, Zoom, Width Shifting, Height Shifting, and Horizontal flip. The results of the data augmentation process are presented in Table III.

TABLE III. Augmented Datasets and Foggy Scene Categories

Foggy Scene	Total
Homogeneous scene	1536
Inhomogeneous scene	1050
Dark scene	1639
Sky scene	2766
Total	**6991**

As shown in Table III, the number of images within each foggy scene type has increased five times from the original set. For instance, the number of homogenous foggy images in the collected dataset was 256 while after applying data augmentation processes, the image number became 1536. Furthermore, other types of foggy images have increased from 175 to 1050 for inhomogeneous foggy images, from 274 to 1639 for dark foggy images, and from 463 to 2766 for foggy sky images. The dataset has increased from 1166 to 6991 foggy images. However, the difference between the original collected and augmented datasets with foggy Scene categories shows in Fig. 4.

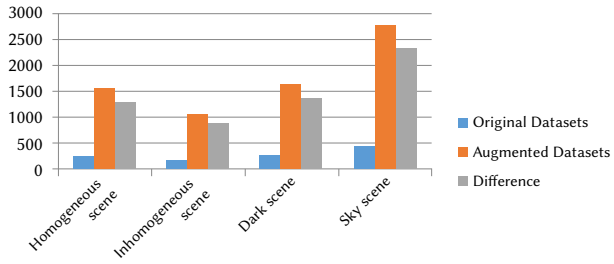


Fig. 4. The difference between original databases collected and augmented datasets with foggy scene categories.

TABLE IV. The Number of Images Used in the Experiment With Each Foggy Scene Class

Class No.	Class Name	Training Set	Validation Set	Testing Set	Total
1	Homogeneous foggy	1108	120	308	1,536
2	Inhomogeneous foggy	760	80	210	1,050
3	Sky foggy	2011	200	555	2,766
4	Dark foggy	1181	130	328	1,639
-	All	5,060	530	1401	6,991

The numbers of images related to each foggy scene category are listed in Table IV. To train and test our network on a multi-fog type dataset, we selected 70% (5060) of sample images evenly from each scene category for training and 10% (530) for validation, and 20 % (1401) as the test set. Adam optimizer was used to fine-tune the parameters while using hyper-parameters with the number of epochs set to 20, batch size = 10, learning rate = 0.0001, momentum = 0.9, and weight decay = 0.0002. AS shown in Table V. To mention, all hyper-parameters have applied for proposed model and state-of-art methods.

TABLE V. Hyper Parameters and Their Values

Hyper-Parameters	Values
Optimization Method	Adam
Momentum	0.9
Weight-Decay	0.0002
Batch Size	10
Activation Function	ReLU
Learning Rate	0.0001
Total No. of Epochs	20
Dropout ration	0.5
# Nodes in the softmax layer	4

The most critical configuration setting in the deep learning model is the learning rate and the number of epochs. To mention, the base for setting the learning rate as 0.0001 is that during experiment configuration, we found that our network with learning rate more than 0.0001 we got only 70% as CAR. When we set the learning rate to 0.0001, the CAR increased to 96%. The same CAR rate (96%) was obtained when we decreased the learning rate to 0. 00001. The more surprising results that at a threshold of more than 0.0001, we found our network cannot detect inhomogeneous foggy scenes instead detect all inhomogeneous foggy images as homogenous, dark, and sky foggy scenes respectively. Thus, a scenario of misleading results has been presented by our network when the learning rate more than 0.0001. The base for selecting the number of epochs as 20 is that during the experiment we found when setting the number of epochs to 5, we got a model with 92 accuracies. Then at 10 and 15 epochs, we found that CAR is the same where have increased to 95%. After that we have increased the number of epochs to 20, we found there is less increase in the CAR rate with a score of 96%. However, to this limit, we found no point in increasing the number of epochs since we achieved a good CAR rate besides the level of CAR has increased with a small rate.

In this work, the performance of four different deep learning models (e.g., ResNet-50, VGG-19, Inception-V3, and Xception) were assessed using the testing set in the foggy dataset. Each model was trained using Adam optimizer using a learning rate policy where the learning rate decreases when learning does not advance for some time. The hyperparameters displayed in Table VI used for training purposes. We use a batch re-balancing strategy for better distribution of each foggy class at a batch level. The proposed foggy classification system was built using the Keras deep learning library with a TensorFlow backend. Based on equations (1, 2,3, and 4) the overall evaluation results for ResNet-50 with other deep learning models are presented in Table VI, where each of CAR, Recall, Precision, and F1-Score are used as indicators for comparison scenario. In the multiclass problems, the calculation is performed as the average of each per-class metric [58]. The best performance has been presented by each of ResNet-50 and VGG-16 models. However, as one can see that comparing with the other three deep learning models, the best performance was achieved using the ResNet-50 model with CAR of 96 %, Recall of 96%, Precision of 96%, and F1-Score of 96%. This was followed by the VGG-16 model that achieved a higher result than Inception-V3 and Xception models in terms of all evaluation metrics values. Each of Inception-V3 and Xception got the poorest performance values among the four deep learning models. Where the minimum values of evaluation metrics have scored by these models, for instance, the CAR value has not exceeded more than 66% for Xception, and 62% for the Inception-V3 model. Xception is better than Inception-V3, especially in terms of F1-Score value. The ResNet-50 model has adequate results as it is the most massive deep learning structure among other pre-trained models (VGG-16, Inception-V3, and Xception). Since the ResNet-50 model allows the information flow through the network with residual connections, that is, the gradient value does not diminish through backpropagation, and the deepest structures have the best classification performance.

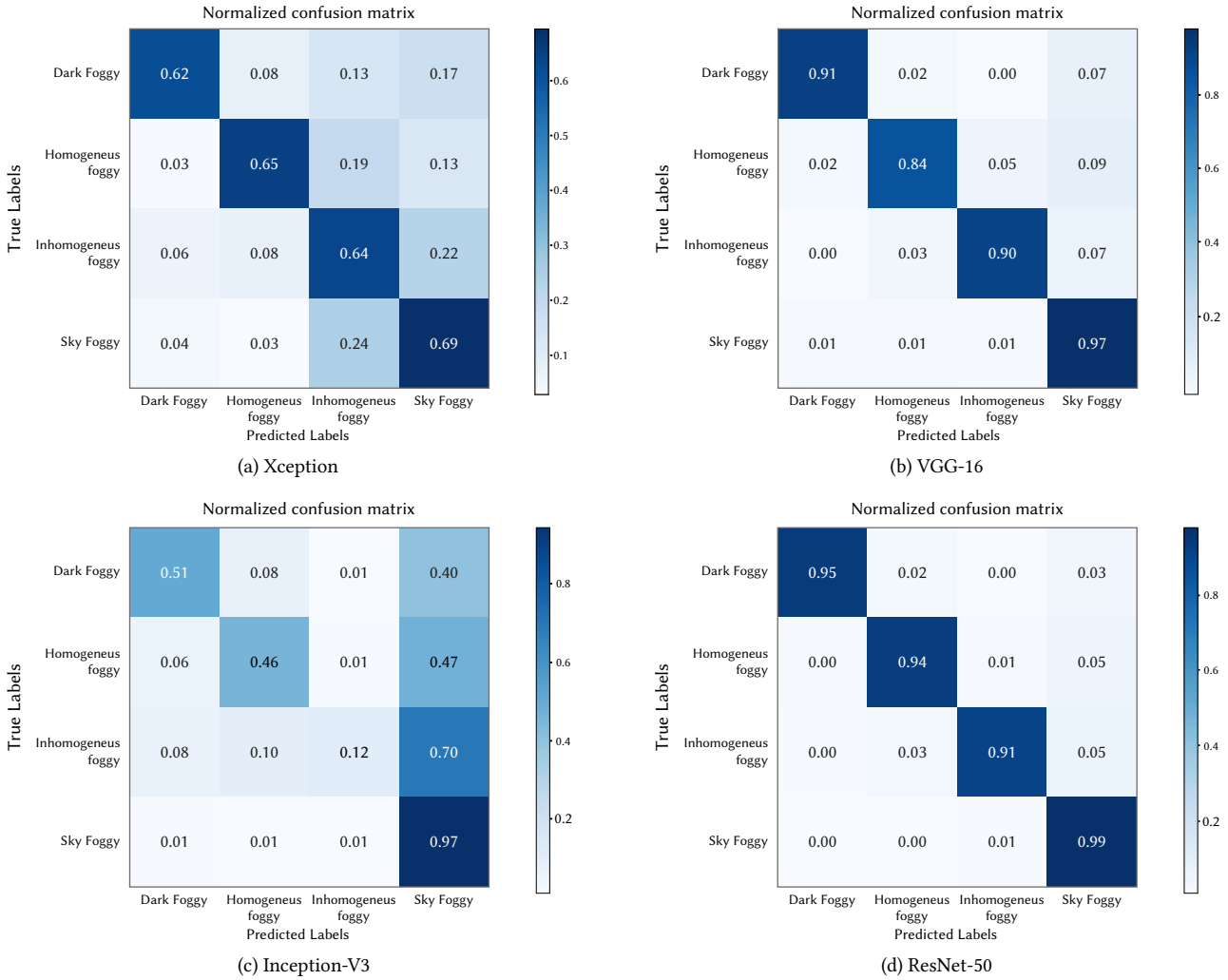
TABLE VI. The Performance Comparison of the Adopted Four Deep Learning Models

Quantitative Measures	Xception	VGG-16	Inception-V3	ResNet-50
CAR	0.66	0.92	0.62	0.96
Recall	0.66	0.92	0.62	0.96
Precision	0.71	0.92	0.67	0.96
F1-Score	0.67	0.91	0.58	0.96

Multi-class classification is prone to imbalance issue which could present compromised performance; therefore, we must highlight the CAR for each class to measure the performance of our model per each foggy scene class. Four confusion matrices with normalized values are presented in Fig. 5.

The ResNet-50 model has the highest classification result per each foggy scene class. This model managed to correctly detect 310 (95% detection rate) out 328 as dark foggy images, 287 (94%) as homogenous foggy images, 193 (91%) as inhomogeneous foggy images, and 536 (99%) as sky foggy images. On the other hand, the VGG-16 model has succeeded to detect 292 (91%) as dark, 274 (84%) as homogenous, 186 (90%) as inhomogeneous, and 529 (97%) as sky foggy images.

The Xception model has a lower rate of detection than the previous two models in all foggy classes, where only 202 (62%) detected as dark, 200 (65%) as homogenous, 134 (64%) as inhomogeneous, and 385 (69%) as sky foggy images. Finally, the lowest detection rate has scored by Inception-V3 in all foggy image classes except for sky foggy class, where this model has presented a significant performance with a detection rate equal to 97% (536 out of 555 foggy sky images); also this model has the same results when comparing with VGG-16 based on detection for sky foggy images. Furthermore, this model has only detected 166 (51%) as dark, 141 (46%) as homogenous, and 25 (12%) as inhomogeneous foggy images. However, based on the results of four deep learning models, overall, the maximum misclassification (low detection) rate has been scored by Inception-V3 in terms of inhomogeneous foggy images where 88% of images are classified incorrectly. On the other hand, the lowest misclassification (highest detection) rate can be seen in the ResNet-50 wherein the foggy sky class only 1% of images



(a) Xception (b) VGG-16

(c) Inception-V3 (d) ResNet-50

Fig. 5. Confusion matrices: (a) Xception (b), VGG-19, (c) Inception-V3, and (d) ResNet-50.

have classified incorrectly. To end, the classification performance is balanced overall classes and in almost all models except Inception-V3 which presented high imbalance classification performance.

V. Discussion

A. Application Scenario

Our foggy image detection model can perform in real-time applications and is suitable as a pre-processing step, for many real-world applications such as basic image defogging, traffic surveillance systems, and driving assistance applications (3-D scene reconstruction in foggy weather, object detection, and recognition, etc.).Our detection model can serve in the driver assistant system as adaptive approach from multi perspectives, first, in the driver assistant system where the car is considered as a non-fixed object that can move through different fog environments, day, night, city, and mountains. For instance, when the car moves on the road in the daytime with fixed fog density in the air, thus the most suitable to detect the foggy scene is the detection algorithm based on the homogenous foggy scene. Second, when the car move inside the city at nighttime and more artificial lights are presented with fog. In this scenario, the most dominant detection algorithm is that based on detecting the dark foggy image. Third, when the car moves on the top of the mountain where the scene is including two aspects. The captured image may contain a wide area of sky in the image. Most of the fog in the mountain area is unevenly distributed; thus, the inhomogeneous foggy scene is presented. Therefore, an adaptive detection algorithm is needed to detect the sky foggy scene besides detecting the inhomogeneous foggy scene whenever presented. To end, our detection model can fulfill the conditions of detection in each one of the previously mentioned cases. Moreover, even if all scenarios are presented together or individually, our algorithm can tackle the issues of detection in different foggy environments; thus, the primary goal of our proposed model is achieved. Furthermore, to provide users with a cost-efficient solution for contrast restoration in driving scenarios, the possibility of our algorithm if can be ported on mobile devices still questionable where more lightweight deep learning architecture is needed for such applications.

1. Comparison With State-Of-The-Art Methods

Since our study classified foggy images based and deep learning model and deep features, these two aspects have used as criteria for selecting the benchmark study.

According to the related works section, we found that the most relevant study to our work was study [22]. However, this study has tested only binary classification where the model classified images as foggy and non-foggy type. For this reason, we implemented the same model based on our dataset where the model classifies images to inhomogeneous, homogeneous, dark, and sky foggy image. Table VII show the comparison with state-of-the-art method based on well know evaluation metrics.

TABLE VII. Comparison With State-of-the-art Methods

Quantitative Measures	Study [22]	Proposed model
CAR	0.67	0.96
Recall	0.67	0.96
Precision	0.71	0.96
F1-Score	0.64	0.96

Table VII showed that proposed model outperformed the benchmark study in all quantitative measures. Thus, the proposed study has presented very efficient performance compared with benchmarked study.

2. Limitations

- Some images are confused with other images such as sky images with homogenous and inhomogeneous where some of these images may contain sky area but with different ranges so maybe the model misclassified images (create overlapping scenario), in other words, classify homogeneous and inhomogeneous images as sky ones. So, with image segmentation technique could solve such issue and define the region of interests in the homogenous and inhomogeneous foggy scene.
- When we configured our model with 0.001 learning rate, we found that our model cannot detect homogenous foggy images at all and have misleading results. Where 194 out of 308 homogenous foggy images were classified as sky foggy images,78 images classified as dark foggy, and 36 images as inhomogeneous foggy. Besides the same learning rate, the CAR rate has been achieved was only 70%. While less than 0.001 we have achieved good CAR beside there are no misleading results.
- More real datasets based on four foggy scenes (inhomogeneous, homogenous, sky, and dark) are needed to build rather than only depend on the image augmentation techniques because even with advantages of this techniques it provides geometrical transformations for the images but unfortunately this does not reflect different fog characteristics in the foggy images which are the base for the scene complexity.

VI. Conclusions

The main contributions of this paper lie in two folds. First, the development of a new detection model for multi-fog scenes based on a deep learning approach. Second, a collected dataset of multi-fog images based on publicly available datasets is presented. A total of 1166 different foggy images are collected from different resources. To provide an efficient training process by tackle overfitting and imbalanced dataset issues, the same set has increased to 6991 foggy images by using data augmentation techniques. More pre-processing procedures have been applied to the datasets. The 6991 images are the basis for the training, validation, and testing of the proposed model.

Our proposed method has successfully detected the foggy images with different fog types, including inhomogeneous, homogeneous, dark, and sky foggy scene. The processes and steps of the proposed detection model were described. The development of the proposed deep learning detection model was formed based on ResNet-50 architecture. To verify the efficiency of the proposed model, an evaluation experiment has been conducted based on different measurements as well as within different deep learning models. The results confirm that:

(1) Comparing with the other three deep learning models; the best performance was achieved using the ResNet-50 model with CAR of 96 %, Recall of 96%, Precision of 96%, and F1-Score of 96%. This was followed by VGG-16 model that achieved a higher result than the Inception-V3 and Xception models in terms of all evaluation metrics values.

(2) The ResNet-50 model has the highest classification result per each foggy scene class. This model managed to correctly detect 310 (95% detection rate) out 328 as dark foggy images, 287 (94%) as homogenous foggy images, 193 (91%) as inhomogeneous foggy images, and 536 (99%) as sky foggy images.

(3) Based on the results of four deep learning models, overall, the maximum misclassification (low detection) rate has been scored by Inception-V3 in terms of inhomogeneous foggy images where 88% of images are classified incorrectly. On the other hand, the lowest misclassification rate can be seen in the ResNet-50 wherein the foggy sky class only 1% of images have classified incorrectly. This is followed by 5% as a misclassification rate for dark, 6% for homogenous, and 8% for inhomogeneous.

Acknowledgment

This work is funded by the projects SP2022/18 and SP2022/34, assigned to VSB-Technical University of Ostrava, the Ministry of Education, Youth and Sports in the Czech Republic.

References

[1] O. Iparraguirre, A. Amundarain, A. Brazalez, and D. Borro, "Sensors on the move: Onboard camera-based real-time traffic alerts paving the way for cooperative roads," Sensors, vol. 21, no. 4, p. 1254, 2021, doi:10.3390/s21041254.

[2] A. Ronen, E. Agassi, and O. Yaron, "Sensing with polarized lidar in degraded visibility conditions due to fog and low clouds," Sensors, vol. 21, no. 7, p. 2510, 2021, doi:10.3390/s21072510

[3] Z. Liu, Y. He, C. Wang, R. Song, "Analysis of the Influence of Foggy Weather Environment on the Detection Effect of Machine Vision Obstacles," Sensors, vol. 20, no. 2, p. 349, 2020, doi: 10.3390/s20020349

[4] G. Broughton, F. Majer, T. Rouček, Y. Ruichek, Z. Yan, and T. Krajník, "Learning to see through the haze: Multi-sensor learning-fusion system for vulnerable traffic participant detection in fog," Robotics and Autonomous Systems, vol. 136, p. 103687, 2021 doi:10.1016/j.robot.2020.103687

[5] D. Nair and P. Sankaran, "Color image dehazing using surround filter and dark channel prior," Journal of Visual Communication and Image Representation, vol. 50, pp. 9-15, 2018.

[6] Y. Xu, J. Wen, L. Fei, and Z. Zhang, "Review of Video and Image Defogging Algorithms and Related Studies on Image Restoration and Enhancement," IEEE Access, vol. 4, pp. 165-188, 2016.

[7] S. G. Narasimhan and S. K. Nayar, "Vision and the atmosphere," International Journal of Computer Vision, vol. 48, no. 3, pp. 233-254, 2002.

[8] K. H. Abdulkareem, et al., "A Novel Multi-Perspective Benchmarking Framework for Selecting Image Dehazing Intelligent Algorithms Based on BWM and Group VIKOR Techniques," International Journal of Information Technology & Decision Making, Vol. 19, no. 3, pp. 909-957, 2020.

[9] K. H. Abdulkareem et al., "A new standardisation and selection framework for real-time image dehazing algorithms from multi-foggy scenes based on fuzzy Delphi and hybrid multi-criteria decision analysis methods," Neural Computing and Applications, Vol. 33, no. 4, pp. 1029-1054, 2021.

[10] W. Wang and X. Yuan, "Recent advances in image dehazing," IEEE/CAA Journal of Automatica Sinica, vol. 4, no. 3, pp. 410-436, 2017.

[11] C. Qu, D.-Y. Bi, P. Sui, A.-N. Chao, and Y.-F. Wang, "Robust Dehaze Algorithm for Degraded Image of CMOS Image Sensors," Sensors, vol. 17, no. 10, p. 2175, 2017.

[12] J.-M. Guo, J.-y. Syue, V. R. Radzicki, and H. Lee, "An efficient fusion-based defogging," IEEE Transactions on Image Processing, vol. 26, no. 9, pp. 4217-4228, 2017.

[13] N. Baig, M. M. Riaz, A. Ghafoor, and A. M. Siddiqui, "Image dehazing using quadtree decomposition and entropy-based contextual regularization," IEEE Signal Processing Letters, vol. 23, no. 6, pp. 853-857, 2016.

[14] K. Kim, S. Kim, and K.-S. Kim, "Effective image enhancement techniques for fog-affected indoor and outdoor images," IET Image Processing, vol. 12, no. 4, pp. 465-471, 2017.

[15] Y. Wang and P. Yuen, "Image dehazing based on partitioning reconstruction and entropy-based alternating fast-weighted guided filters," Optical Engineering, vol. 56, no. 5, p. 053111, 2017.

[16] A. Alajarmeh, R. A. Salam, K. Abdulrahim, M. F. Marhusin, A. A. Zaidan, and B. B. Zaidan, "Real-time framework for image dehazing based on linear transmission and constant-time airlight estimation," Information Sciences, vol. 436–437, pp. 108-130, 2018.

[17] K. Tang, J. Yang, and J. Wang, "Investigating Haze-Relevant Features in a Learning Framework for Image Dehazing," in 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2995-3002.

[18] Y. Shi and X. Jiang, "Deep quality assessment toward defogged aerial images," Signal Processing: Image Communication, vol. 83, p. 115808, 2020.

[19] C. Qing, Y. Hu, X. Xu, and W. Huang, "Image haze removal using depth-based cluster and self-adaptive parameters," in 2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), 2017, pp. 1070-1075: IEEE.

[20] M. Pavlić, H. Belzner, G. Rigoll, and S. Ilić, "Image based fog detection in vehicles," in 2012 IEEE Intelligent Vehicles Symposium, 2012, pp. 1132-1137: IEEE.

[21] Y. Liu, J. Pan, J. Ren, and Z. Su, "Learning deep priors for image dehazing," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 2492-2500.

[22] L. Guo et al., "Haze Image Classification Method Based on Alexnet Network Transfer Model," in Journal of Physics: Conference Series, 2019, vol. 1176, no. 3, p. 032011: IOP Publishing.

[23] H. Zhang, V. Sindagi, and V. M. Patel, "Multi-scale single image dehazing using perceptual pyramid deep network," in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 902-911.

[24] C. Hernandez-Olivan, J. R. Beltran, and D. J.a.p.a. Diaz-Guerra, "Music Boundary Detection using Convolutional Neural Networks: A comparative analysis of combined input features," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 6, pp. 208-214, 2021, https://doi.org/10.9781/ijimai.2021.01.003.

[25] M.I. Khattak, et al., "Automated detection of COVID-19 using chest x-ray images and CT scans through multilayer-spatial convolutional neural networks," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 6, pp. 15-24, 2021. https://doi.org/10.9781/ijimai.2021.04.002.

[26] Y. Chen, J. Wang, S. Li, and W. Wang, "Multi-feature based Foggy Image Classification," in IOP Conference Series: Earth and Environmental Science, 2019, vol. 234, no. 1, p. 012089: IOP Publishing.

[27] M. Pavlic, G. Rigoll, and S. Ilic, "Classification of images in fog and fog-free scenes for use in vehicles," in 2013 IEEE Intelligent Vehicles Symposium (IV), 2013, pp. 481-486: IEEE.

[28] Y. Xu, J. Wen, L. K. Fei, and Z. Zhang, "Review of Video and Image Defogging Algorithms and Related Studies on Image Restoration and Enhancement," IEEE Access, vol. 4, pp. 165-188, 2016.

[29] H. Shi, Q. Wang, and L. Xie, "A Method of Automatic Detection of Fog Image Based on SVM Classification," Revista de la Facultad de Ingeniería, vol. 31, no. 9, pp. 211-218, 2016.

[30] X. Yu, C. Xiao, M. Deng, and L. Peng, "A classification algorithm to distinguish image as haze or non-haze," in 2011 Sixth International Conference on Image and Graphics, 2011, pp. 286-289: IEEE.

[31] J.-P. Tarel, N. Hautiere, A. Cord, D. Gruyer, and H. Halmaoui, "Improved visibility of road scene images under heterogeneous fog," in 2010 IEEE Intelligent Vehicles Symposium, 2010, pp. 478-485: IEEE.

[32] Y. Li, R. T. Tan, and M. S. Brown, "Nighttime haze removal with glow and multiple light colors," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 226-234.

[33] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," IEEE transactions on pattern analysis and machine intelligence, vol. 33, no. 12, pp. 2341-2353, 2011.

[34] D. Berman, T. Treibitz, and S. Avidan, "Air-light estimation using haze-lines," in 2017 IEEE International Conference on Computational Photography (ICCP), 2017, pp. 1-9: IEEE.

[35] Y. Zhang, G. Sun, Q. Ren, and D. Zhao, "Foggy images classification based on features extraction and SVM," in 2013 International Conference on Software Engineering and Computer Science, 2013: Atlantis Press.

[36] S. Shrivastava, R. K. Thakur, and P. Tokas, "Classification of hazy and non-hazy images," in 2017 International Conference on Recent Innovations in Signal processing and Embedded Systems (RISE), 2017, pp. 148-152: IEEE.

[37] Y. Liu, H. Li, and M. Wang, "Single image dehazing via large sky region segmentation and multiscale opening dark channel model," IEEE Access, vol. 5, pp. 8890-8903, 2017.

[38] Y. Song, H. Luo, J. Ma, B. Hui, and Z. Chang, "Sky detection in hazy image," Sensors, vol. 18, no. 4, p. 1060, 2018.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778.

[40] Z. Chen, Z. Xie, W. Zhang, and X. Xu, "ResNet and Model Fusion for Automatic Spoofing Detection," in INTERSPEECH, 2017, pp. 102-106.

[41] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," 2015. Arxiv:1505.00387

[42] R. Fattal, "Dehazing using color-lines," ACM transactions on graphics (TOG), vol. 34, no. 1, p. 13, 2014.

[43] C. O. Ancuti, C. Ancuti, R. Timofte, and C. De Vleeschouwer, "O-haze: a dehazing benchmark with real hazy and haze-free outdoor images," in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 754-762.

[44] K. D. Ma, W. T. Liu, and Z. Wang, "Perceptual evaluation of single image dehazing algorithms," in 2015 Ieee International Conference on Image Processing, 2015, pp. 3600-3604.

[45] L. K. Choi, J. You, and A. C. Bovik, "Referenceless prediction of perceptual fog density and perceptual image defogging," IEEE Transactions on Image Processing, vol. 24, no. 11, pp. 3888-3901, 2015.

[46] A. Galdran, J. Vazquez-Corral, D. Pardo, and M. Bertalmio, "Fusion-based variational image dehazing," IEEE Signal Processing Letters, vol. 24, no. 2, pp. 151-155, 2016.

[47] B. Li et al., "Benchmarking Single-Image Dehazing and Beyond," IEEE Transactions on Image Processing, vol. 28, no. 1, pp. 492-505, 2019.

[48] J. Li, G. Li, and H. Fan, "Image dehazing using residual-based deep CNN," IEEE Access, vol. 6, pp. 26831-26842, 2018.

[49] Y. Qu, Y. Chen, J. Huang, and Y. Xie, "Enhanced pix2pix dehazing network," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 8160-8168.

[50] J. Zhang and D. Tao, "FAMED-Net: a fast and accurate multi-scale end-to-end dehazing network," IEEE Transactions on Image Processing, vol. 29, pp. 72-84, 2019.

[51] X. Zhu, X. Xu, and N. Mu, "Saliency Detection Based on the Combination of High-Level Knowledge and Low-Level Cues in Foggy Images," Entropy, vol. 21, no. 4, p. 374, 2019.

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

[53] Abayomi-Alli, O. O., Damasevicius, R., Maskeliunas, R., & Abayomi-Alli, A. (2020). "BiLSTM with data augmentation using interpolation methods to improve early detection of parkinson disease," Proceedings of the 2020 Federated Conference on Computer Science and Information Systems, FedCSIS 2020, 371-380. doi:10.15439/2020F188

[54] Abayomi-Alli, O. O., Damaševičius, R., Wieczorek, M., & Woźniak, M. (2020). "Data augmentation using principal component resampling for image recognition by deep learning," In Artificial Intelligence and Soft Computing (pp. 39–48). Springer International Publishing. doi:10.1007/978-3-030-61534-5_4

[55] Arif, Zainab Hussein, et al. "Comprehensive Review of machine Learning (ML) in Image Defogging: Taxonomy of Concepts, Scenes, Feature Extraction, and Classification techniques," IET Image Processing (2021)

[56] K. H. Abdulkareem, et al., "Mapping and Deep Analysis of Image Dehazing: Coherent Taxonomy, Datasets, Open Challenges, Motivations, and Recommendations," International Journal of Interactive Multimedia and Artificial Intelligence, Vol. 7, no. 2, pp. 172-198, 2021.

[57] Al-Mhiqani, Mohammed Nasser, et al. "New insider threat detection method based on recurrent neural networks," Indonesian Journal of Electrical Engineering and Computer Science, vol. 17, no. 3, pp. 1474-1479, 2020.

[58] M. Hossin, and D. M. Sulaiman, "A review on evaluation metrics for data classification evaluations," International Journal of Data Mining & Knowledge Management Process, vol. 5, no. 2, pp. 1-11, 2015.

Zainab Hussein Arif

Zainab Hussein Arif received the B.Sc. degree in computer science and information technology, in 2016 from University of Qadisiyah,.She obtained master degree Universiti Tenaga Nasional (Uniten) in 2021 in computer science and information technology. Her research interests include data science and biomedical computing. Link:https://scholar.google.com/citations?user=ecBErWQAAAAJ&hl=en&authuser=1

Moamin A. Mahmoud

Moamin A. Mahmoud received the bachelor's degree in mathematics from the College of Mathematics and Computer Science, University of Mosul, Iraq, in 2007, the Master of Information Technology degree from the College of Graduate Studies, Universiti Tenaga Nasional (UNITEN), Malaysia, in 2010, and the Ph.D. degree in information and communication technology from UNITEN, in 2013. Since 2014, he has been with the Department of Software Engineering, Universiti Tenaga Nasional, as a Senior Lecturer. His current research interests include artificial intelligence, distributed and autonomous systems, complex adaptive systems, and the IoT-based smart systems.

Karrar Hameed Abdulk

Karrar Hameed Abdulkareem received the B.S. degree in computer science (Artificial Intelligence) from the University of Technology, Iraq, in 2007, and the M.S. degree in computer science (Internetworking Technology) from the Universiti Teknikal Malaysia Melaka (UTeM), Malaysia, in 2016. He Obtained Ph.D. degree in Computer Science and Information Technology from Universiti Tun Hussein Onn Malaysia (UTHM), Malaysia. He has produced more than 55 articles into different ISI Web of Science journals, such as IEEE internet of things, Journal of King Saud University - Computer and Information Sciences, Computer methods and programs in biomedicine (Elsevier), Neural Computing and Applications (Springer), IEEE Access, Journal of infection and public health (Elsevier), International Journal of Information Technology & Decision Making (World Scientific), Computers, Materials & Continua (Tech Science Press), Soft Computing (Springer), and Sensors (MDPI). He has a total number of citations over 1666 (Google Scholar H-Index = 23). He has collaborated with many researchers over international countries. He has served as a reviewer for more than 28 international journals. His research area includes Multi-Criteria Decision Making, Artificial Intelligence, Data Science, and Fog Computing.

Seifedine Kadry

Seifedine Kadry (Senior Member, IEEE) received the bachelor's degree from Lebanese University, in 1999, the dual M.S. degree from Reims University, France, in 2002, and EPFL, Lausanne, the Ph.D. degree from Blaise Pascal University, France, in 2007, and the H.D.R. degree from Rouen University, in 2017. His current research interests include data science, education using technology, system prognostics, stochastic systems, and probability and reliability analysis. He is an ABET Program Evaluator of computing and an ABET Program Evaluator of Engineering Tech.

Mazin Abed Mohammed

Mazin Abed Mohammed received the B.Sc. degree in computer science from the University of Anbar, Iraq, in 2008, the M.Sc. degree in information technology from UNITEN, Malaysia, in 2011, and the Ph.D. degree in information technology from UTeM, Malaysia, in 2019. He is currently a Lecturer with the College of Computer Science and Information Technology, University of Anbar, Iraq. His research interests include artificial intelligence, biomedical computing, and optimization.

Mohammed N. Al-Mhiqani

Mohammed N. Al-Mhiqani received his BSc in Computer Science (Computer Networking) in 2014, his MSc in Computer Science (Internetworking Technology) from Universiti Teknikal Malaysia Melaka (UTeM) in 2015, and his PhD in cybersecurity and artificial intelligence from UTeM in 2022. He joined the School of Computing and Mathematics at Keele University in 2022 as a postdoctoral research associate. His research interests include cybersecurity, cyber physical systems security, insider threats, artificial intelligence, and health informatics.

Alaa S. Al-Wais

Alaa S. Al-WaisY is a doctor of philosophy in computer science. He received his BSc in 2009 and MSc in 2011 in computer science from Al-Anbar University. In 2018, he received PhD in digital imaging and visualization from the University of Bradford. His research interests include pattern recognition, image processing, computer vision, medical imaging, designing and implementing unimodal and multimodal biometric systems.

Jan Nedoma

Jan Nedoma (Senior Member, IEEE) is currently an Associate Professor and the Head of the Optoelectronics Laboratory with the Department of Telecommunications, Faculty of Electrical Engineering and Computer Science, Technical University of Ostrava. He is a member of the scientific council, a member of doctoral, habilitation, and professors' committees, and a Guarantor of bachelor's study programs at Faculty of Electrical Engineering and Computer Science, Technical University of Ostrava. During his scientific career, he was the Leader or a co-investigator of more than 25 projects and has more than 175 journal articles and conference papers in his research areas. He holds ten valid Czech patents. His research interests include optical communications, optical atmospheric communications, optoelectronics, optical measurements, measurements in telecommunication technology, signal processing, fiber-optic sensors, and biomedical engineering.

Brain Tumor Segmentation and Identification Using Particle Imperialist Deep Convolutional Neural Network in MRI Images

Maahi Amit Khemchandani*, Shivajirao Manikrao Jadhav, B. R. Iyer

Dr. Babasaheb Ambedkar Technological University Vidyavihar, Lonere, Maharashtra 402103 (India)

Received 7 October 2020 | Accepted 24 December 2021 | Published 17 October 2022

Abstract

For the past few years, segmentation for medical applications using Magnetic Resonance (MR) images is concentrated. Segmentation of Brain tumors using MRI paves an effective platform to plan the treatment and diagnosis of tumors. Thus, segmentation is necessary to be improved, for a novel framework. The Particle Imperialist Deep Convolutional Neural Network (PI-Deep CNN) suggested framework is intended to address the problems with segmenting and categorizing the brain tumor. Using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) Algorithm, the input MRI brain image is segmented, and then features are extracted using the Scatter Local Neighborhood Structure (SLNS) descriptor. Combining the scattering transform and the Local Neighborhood Structure (LNS) descriptor yields the proposed descriptor. A suggested Particle Imperialist algorithm-trained Deep CNN is then used to achieve the tumor-level classification. Different levels of the tumor are classified by the classifier, including Normal without tumor, Abnormal, Malignant tumor, and Non-malignant tumor. The cell is identified as a tumor cell and is subjected to additional diagnostics, with the exception of the normal cells that are tumor-free. The proposed method obtained a maximum accuracy of 0.965 during the experimentation utilizing the BRATS database and performance measures.

Keywords

Brain Tumor Segmentation, Deep Belief Network, LNS Descriptor, Scattering Transform, Tumor Level.

DOI: 10.9781/ijimai.2022.10.006

I. Introduction

The structure of the human brain is highly complex, which is bound inside the skull such that the diagnosis of the diseases becomes a hectic phenomenon. Brain tumors are abnormally growing clumps of brain cells that are visible [1]. Gliomas are a prevalent type of brain tumor that have significant fatality rates [2], [3]. After the diagnosis of Gliomas, the survival of patients is not more than 14 months [4]. A brain tumor diagnosis is made through surgery, radiotherapy, or a combination of all [5], [6], which is a hectic process. The tumors can be of two types benignor malignant [7]. For the benign tumor, the tumor mass does not affect the nearby healthy cells, and they are non-cancerous, whereas the malignant tumors develop as a cancerous mass leading to death when untreated [8], [1]. The modalities of X-ray, CT, Ultrasonography, and MRI are used to collect information on brain tumors so that clinicians can better understand the tumors' textural properties and choose the best course of treatment. The association of the manual analyzing and segmenting the brain tumor images can lead to computational segmentation and classification [9].

MRI offers a better imaging mechanism and assists in accessing the gliomas in such a way to obtain the MRI sequences that yield effective information [2], [6]. Using the diverse features of brain tumors, MRI renders viable information regarding the tumor of various tissues[10], [11]. With the advances in medical imaging technologies, tumor segmentation methods [12], [13] assure the automatic, repeatable, and accurate tumor segmentation algorithm rendering variability both in inter-and intra-rater, with difficulty to reproduce [14], [15]. There are numerous techniques for improving the accuracy and dependability of brain tumor segmentation in the literature [16]. Automatic segmentation [17] helps in better planning of the diagnostic followed by the surgical operations [18] because segmentation is a hectic and demanding component. The manual segmentation for planning in the clinical routines [19] results in human errors and it is time-consuming. Moreover, the radiologists require additional knowledge regarding the pathology and inflexible nature [20], [21] of the existing algorithms led to the real clinical practice because of the development of the non-reliable method using the ill-trained information and parameters that are inflexible in various datasets [10].

The segmentation method uses the brain tumor images for extracting the abnormal tissues for learning the shape and the growth with time. At the same time, tumor delineation using manual methods is a tedious process as a large number of incremental medical data is available without considering the time and visual perception of the health professionals. The above demerits are tackled using computational methods as they are highly accurate and faster [16]. For the last decade, several brain tumor segmentation methods, which are grouped into two categories, semi-automatic and automatic, are available. The first group of methods includes level-set models [22], active contour models [23], and Tumor Cut (TC) model [24] and these models

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

operate independently without the need for humans, but the results are not satisfactory [15]. The automatic approaches for segmenting brain tumors make use of manually created characteristics, and the models incorporate the standard machine learning process [25]. An alternative approach to model the task-adapted feature understands the hierarchy based on the highly complex features straight from the in-domain data. On the other hand, deep neural networks [26] render better learning including feature hierarchies [25].

The paper's main aim engages in classifying the levels of the tumor using MRI brain images through a novel framework, and it consists of three steps. DBSCAN is initially used to segment the picture of the brain tumor. The SLNS descriptor, which combines the scattering transform and LNS descriptors, is then used to the feature extraction. The suggested PI Deep CNN, which uses feature vectors created using the histogram of the output from the SLNS descriptor and the grid-based shape features, is then used to advance the classification of brain tumors. The created descriptor effectively renders the extremely durable texture features for categorization. The classifier focuses at detecting cancers and its level is divided into four classes, such as Normal without tumor, Abnormal, Malignant tumor, and Non-malignant tumor. Deep CNN is optimally tuned using the proposed PI algorithm.

The contributions:

proposed classifier for the particle imperialist deep convolutional neural network (PI-deep CNN): The classifier's goal is to categorize the brain tumor using the features that were derived from the MRI input image segments in order to identify the different levels of the tumor. The suggested PI algorithm, which combines PSO and the Imperialist colony algorithm, is used to optimize the tuning of the DEEP CNN classifier.

Scatter Local Neighborhood Structure (SLNS) description that has been proposed: The suggested descriptor, SLNS, which combines the LNS and scattering transform descriptors, extracts texture information in a way that the robust features support accurate tumor categorization.

The remainder of this essay is structured as follows: The methods for segmenting and categorizing brain tumors that are currently used in the literature are reviewed in Section 2. The automatic segmentation utilizing DBSCAN for tumor-level categorization using a deep learning system is described in Section 3. The results and comments are explained in Section 4, and the paper is wrapped up in Section 5.

II. Motivation

The review of the existing methods for segmentation is demonstrated in this section that deliberates the need for the new model for effective segmentation. The methods developed by various authors and their demerits associated with tumor detection are discussed. Finally, the challenges of the methods are described laying out a smooth pathway to propose an effective brain tumor segmentation and classification method.

A. Related Works

This section is an overview of the literature on several techniques for MRI image segmentation of brain tumors. Based on segmentation of brain tumors, these research papers are selected and evaluated in accordance with the most recent years of publication. As a result, the following 8 research publications are chosen for reviews:

Convolutional Neural Networks (CNN) were utilized by Sergio Pereira et al. [6] to automatically segregate the brain tumors from the MRI images. Although the strategy reduced computing time, it performed poorly when there were several feature maps present. Elisee Ilunga-Mbuyamba and colleagues [16] created an alternative Active Contour Model (ACM) that is based on the Multi-population Cuckoo Search (MCSS) algorithm, which solved the segment's energy reduction issue. It rendered better accuracy with minimal computational times, but the method was more expensive than the existing technique. A total of five PCA algorithms were used by Irem Ersoz Kaya et al. [9] that aimed at dimensionality reduction but suffered from overfitting problems. Chaiyanan Sompong et al. [10] used a framework that improved brain tumor segmentation. The presence of the ambiguous tumor boundaries was tackled using the Gray-Level Co-Occurrence Matrix-based Cellular Automata (GLCM-CA) that transformed the original image as the target featured image, which is highly complex. Lubna Farhi et al. [1] designed an adaptive stochastic segmentation algorithm to model the energy-based stochastic segmentation that possessed a highly flexible topology with high speed and segmentation accuracy. It involved minimum iterations and less computation time with the help of the internal and external forces, but the invariability of the contour radius minimized the requirement of user intervention. A sparse representation-based algorithm developed by Yuhong Li et al. [15] offered better likelihood estimation but did not use deep learning and failed to render accurate results. Yuhong Li et al. [25] modeled an automatic method using the Deep Neural Networks (DNNs) that was fast and accurate in segments, but the presence of a large number of the outliers was a drawback. Elisee Ilunga-Mbuyamba et al. [27] used Localized Active Contour with Background Intensity Compensation (LACM-BIC) method for determining the abnormal tissue, but the assumption of the method regarding the tumor mass was a real failure.

B. Challenges

- GLCM-CA [10] offered better results inthe case of the regions of white matter that possess similar intensities of tumor region. Additionally, the patch weight distance based on the tumor voxels is exceedingly complicated to compute using the Improved Tumor Cut method (ITC).
- In an adaptive stochastic segmentation algorithm [1], the lack of the possibility to automatically change the contour radius insists on the need for human intervention. Moreover, the algorithm is inapplicable for a 3Dform of images.
- The existence of the invariability of the inter and intra intensities among the training and test data poses a huge challenge regarding the segmentation of the enhancing and core parts. Mainly these parts are not clear in the BRATS-HG0117 and HG0307 [15].
- The LACM-BIC framework [27] assumes that the images are initially grouped as tumor mass. In the case of the absence of the classified datasets, the segmentation results in inaccurate results.
- The segmentation is highly complex and time-consuming and does not focus on highly robust and accurate methods that need improving the computational speed in real-time applications.
- CNN [6] for segmenting the brain tumor suffered from the computational complexity and needed more training data.

III. Automatic Segmentation Using DBSCAN for Tumor-Level Classification Using Deep Learning Algorithm

The automatic classification of brain tumor reduces the time spend on the MRI images for segmenting and detecting the affected regions of the brain. Moreover, the detection of the affected region through traditional vision-based techniques is a tedious process. Therefore, the suggested method of PI-based Deep CNN is utilized to categorize the affected region utilizing the histogram and shape aspects of the segments, making it easier and more automated to detect brain tumors. The DBSCAN clustering technique is used to advance the segmentation, which groups the MRI brain image slices into tumor and non-tumor regions. Using the suggested grid-

based SLNS descriptor, which combines the effects of the scattering transform and the LNS descriptor, the characteristics of the segments are retrieved. The afflicted areas of the brain can be identified using Deep CNN, an artificial classifier that was developed using the Imperialist Competitive Algorithm. The architecture of the tumor-level classification is depicted in Fig. 1.

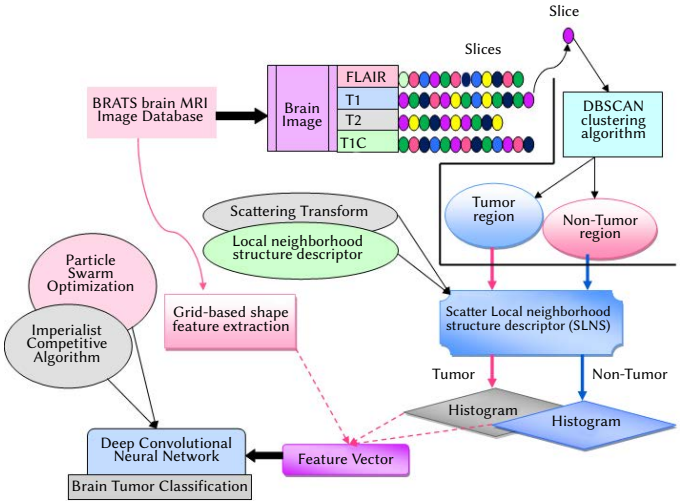


Fig. 1. Schematic diagram of the tumor classification.

For classification, MRI brain images are used that possess four modalities, FLAIR, T1, T2, and T1C, and let us denote the database as, D. Each modality possesses l_a slice and is subjected to segmentation using the DBSCAN clustering algorithm.

A. Formation of Segments Using DBSCAN Clustering

The goal of segmentation is to identify the affected regions in the MRI brain image using their modalities. The segments are generated from the image in such a way that these segments share identical properties, like contrast, color, boundary, and texture. The segmentation is done using the DBSCAN clustering algorithm [28], a density-based algorithm, which offers better segmentation accuracy through geometric constraints. The segmentation algorithm exhibits a faster generation of the super-pixels and renders improved boundary adherence in such a way that the pixels exhibiting homogeneous appearance (color and shape) form compact clusters. During clustering, two sets are initialized, candidate and unlabeled sets along with it.The top-down image pixel is named as the seed which is grouped with the labeled set. Thus, there are a total of three types of pixels, such as labeled, unlabeled, and seed. Initially, determine the four unlabeled pixels corresponding to the neighboring pixel of the labeledpixel such that the distance from the unlabelled pixel to the seed is computed. Whenever the distance is found to be less than the threshold, the unlabeled set is added to the candidate cluster that replaces the labeled set in such a way that the unlabelled pixel obtains the label of the seed. The stages are continued until the halting requirement, which is dependent on the total number of pixels and the empty set, is met. The threshold is calculated by dividing the image size by the total number of user-specified superpixels. Below is provided the DBSCAN clustering algorithm's pseudo-code.

The DBSCAN algorithm forms the segments using the input brain tumor image that is given in (1),

$$n = \{n_1, n_2, .n_3\} \tag{1}$$

where n signifies the segments that represent the tumor and the non-tumor region.

Algorithm for DBSCAN Clustering

DBSCAN clustering

Input: Image S, superpixel ρ, threshold T, labeled set L, Candidate set C
Output: Label of ρ, $L(\rho)$
Read the input image S
  Set the initial pixel as '0.'
  For a pixel with new label
  do
    Determine seed s such that $s \in L$
      While $L = \{\}$ or pixels in $\rho > T$ **do**
        For individual pixel i in L do
          For individual k pixel around i pixel do
  Compute the clustering distance $\partial_1^s(i,j)$ with seed s and pixel i
            If $\partial_1^s(i,j) < \delta$ then
              Set $k \in C$
            End if
          Enf For
        End For
          Set $L = C$
      End while
  End For

B. Extraction of the Features From the Segments Using the Proposed Scatter Local Neighborhood Structure Descriptor (SLNS)

Feature extraction ensures accurate and effective classification through the dimensionally-reduced significant features of the segments. The features used include the SLNS features and the shape features that are obtained using the proposed SLNS descriptor and the grid-based shape descriptor. The segments describe the tumor and the non-tumor region of the image in such a way that the texture features from both the regions are extracted separately using the proposed SLNS and grid-based descriptors.

The suggested SLNS descriptor, which combines the work of the scattering wavelet transform [29] and LNS descriptor [30], is used to extract features from the segments extracted using the DBSCAN clustering algorithm. The importance of SLNS lies in its ability to successfully extract textural information from the segments in order to facilitate the precise classification of the tumor level. Each of the segments of the image is subjected to the proposed SLNS descriptor, which yields the features by multiplying the output from scattering transform and LNS descriptor. The steps of the proposed SLNS descriptor are deliberated below.

Step 1: Application of scattering transform: The Morlet wavelet of different scales and orientation is extracted using the Scattering Transform (ST) [31] from separate segments to produce a highly resilient and locally invariant feature. To begin with, the segments' non-linear invariants are obtained using the modulus and average pooling functions. The ST steps are:

i) Preserving the image using local affine transformation: The local affine transformation, which is carried out by convoluting the individual image segments, and the low-pass filter, which functions based on the scaling factor, are used to safeguard the image against deformations. The components with high frequency are eliminated by the local affine transformation.

ii) Morlet filters for capturingthe high-frequency components: By averaging the coefficients generated from wavelet modulus, which are the unique results of the Morletor bandpass and the average filters, the high-frequency components are caught.

iii) Generation of the scattering coefficients: The wavelet modulus transformations are applied to the coefficients of wavelet modulus to produce the scattering coefficients. Convoluting the wavelet modulus coefficients with the generated wavelets of different scales and orientations yields the high-order scattering coefficients. Higher-order coefficients are required because they provide robustfeatures that are extremely stable and invariant locally. The features extracted from the segments using the scattering transform are denoted as, K^{scat}.

Step 2: Application of LNS descriptor:The segmented region is fed to the LNS descriptor [30] such that the individual pixel is thresholded using the neighborhood pixels of the image with the reference value. The reference value is fixed using the center pixel of the image and the mean absolute deviation computed using the difference in the neighboring pixels. The threshold is determined using equation (2).

$$H = G_c + \frac{1}{J}\sum_{r=0}^{J-1}\left|d_r - \frac{1}{J}\sum_{r=0}^{J-1}|d_r|\right| \tag{2}$$

$$d_r = \{G_r - G_c, r = (0,\ldots,J-1) \tag{3}$$

where G_n represents the gray values J that are spaced equally in the neighboring pixels and n vary between 0 and $J-1$.The gray value of the center pixel is denoted as G_c. d_r represents the difference in the local neighborhood and it is calculated using equation (3).

The mean absolute deviation offers robust features and renders the statistical features. The feature vector is obtained through the subtraction of the new threshold from the gray values in the circular neighborhood as given in (4).

$$W = \upsilon\left(G_0 - H, G_1 - H, \ldots, G_{J-1} - H\right) \tag{4}$$

where υ belongs to the joint distribution of J differences and W symbolizes the texture patterns obtained from the neighborhood pixels, whih is computed using (5). The representation is given in binary such that the gray values of the pixels exceeding the threshold are marked as '1' and the gray values lying below the threshold are filled as '0' as given in (6).

$$W = \upsilon\left(v[G_0 - H], v[G_1 - H], \ldots, v\left[G_{J-1} - H\right]\right) \tag{5}$$

$$v(x) = \begin{cases} 1 & ; x \geq 0 \\ 0 & ; x < 0 \end{cases} \tag{6}$$

where $v(x)$ specifies the sign function. To extract the pattern, the binomial 2^J is multiplied with the individual $v(x)$, and all the individuals are added together to represent the texture pattern. Thus, the LNS feature corresponding to each segment of the image is denoted as, K^{LNS}.

Step 3: SLNS feature generation: The third step is integrating the SLNS features using the proposed SLNS descriptor that is developed through the integration of the scattering features and LNS features. The generation of the SLNS features is described in (7).

$$K = \left[K^{Scat} \times K^{LNS}\right] \tag{7}$$

where K refers to the SLNS features. The features obtained from the ST and the LNS descriptor are denoted as, K^{Scat} and K^{LNS}, respectively.

Step 4: Histogram-based features using the SLNP-generated output:The histogram features of the SLNP are determined such that the histogram of the SLNP features belonging to the tumor and the non-tumor region enable the robust and accurate classification of the tumor level. The histogram features for the tumor segment and the non-tumor segment are denoted as p_1 and p_2.

3.2.2 Grid-based shape feature: The grid method is employed for extracting the shape features for which initially, the input image is subdivided with the grid lines. The image is scanned on the top-bottom and the left-right approach so that the grid space that covers the space partially or fully is assigned '1', whereas the grid space with no shape is filled with '0'. The shape feature is indicated as, g.

3.2.3 Feature vector: The feature vector represents the features obtained from the segments of the image, and is given in (8).

$$F = \{p_1, p_2, g\} \tag{8}$$

The dimension of the feature vector is denoted as, [1 x 128].

C. Brain Tumor Classification Using PI-Based Deep CNN Classifier

The Deep CNN classification module receives the extracted features from the segments, tumor and non-tumor, and the PI algorithm is used to optimize the classifier's weights. The PI algorithm, which is the modification of PSO [32] with the Imperial Competitive Algorithm [33] in such a way that the PI-Deep CNN classifies the input features corresponding to the input image to derive four classes. The classes include normal, abnormal, malignant, and non-malignant that are graded as four levels. The feature vector includes the histogram features of the tumor region, non-tumor region, and shape features of the input MRI brain image to facilitate the effective classification of the tumor level with better accuracy.

Comparing the PSO method to mathematical algorithms and other heuristic optimization techniques reveals that it is computationally efficient, simple to implement, robust to control factors, and has a straightforward idea.PSO follows the swarm behavior exhibiting easy implementation,and it requires only a few adjustable parameters compared with the Genetic Algorithm (GA). The particles in the search space are initialized randomly,and the optimal solution is derived through the frequent update of the solutions in the individual iterations. PSO follows the bird flocking behavior in such a way that the birds fly in groups in search of food randomly with certain velocities that direct the bird to fly to various locations. The optimal position of the particle is based on minimal error. The exploration abilities of PSO are high improving the optimal convergence faster. Moreover, PSO depends on intelligence and is applicable both in scientific and engineering applications. On the other hand, the Imperial Competitive Algorithm depends on the political-social developments of humans, and the competition among the imperials begins with the advent of the primary imperials. The decision is made based on the power such that the imperial with high power leaves the competition and the one with the highest power becomes the empire. The inclusion of the Imperial Competitive Algorithm increases the global optimum and minimal global solutions that assure the accuracy of classification.

a) Architecture of the Deep Convolutional Neural Network

Convolutional (Conv) layers, pooling (POOL) layers, and a fully connected (FC) layer make up the deep CNN [34], [35], whose structure is shown in Fig. 2. Each layer more accurately carries out its task of extracting the feature maps, sub-sampling, and classification. The pooling layer, which is used to generate the output maps in the FC layer, subsamples the feature maps created using the conv layer. The quantity of conv layers is used to improve classification accuracy by making the feature maps smaller when more conv layers are added.

Convolutional layers: Convolutional layers are the feature maps of neurons that extract patterns from the features of the input image. Using trainable weights, the neuron's receptive fields connect the neuron in the preceding layer to the neuron in the following convolutional layer. The trainable weights are used to convolute the input features and create a feature map, which is then applied to the subsequent layers using a non-linear activation function. Using the

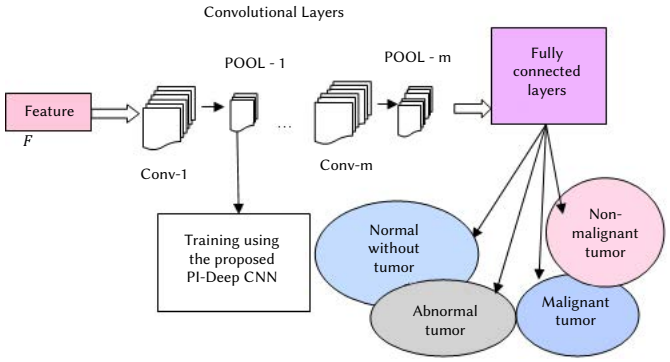


Fig. 2.The architecture of Deep CNN.

same neurons in a single feature map with varied weights in a single conv layer ensures the extraction of the variable features from diverse places. The feature vector serves as the input to the convolution layers of the deep CNN, and the convolution layers are represented in (9).

$$c = \{c_1, c_2, \ldots, c_j, \ldots, c_m\} \tag{9}$$

where m refers to the totalconv layers. The conv layers process the inputs and generate the output and the unit at (y, z) generates the output based on equation (10).

$$(V_u^j)_{y,z} = (b_u^j)_{y,z} + \sum_{q=1}^{f_1^{q-1}} \sum_{\upsilon=-w_1^j}^{w_1^j} \sum_{\kappa=-w_2^j}^{w_2^j} \left(\beta_{u,q}^j\right)_{\upsilon,\kappa} * \left(V_q^{j-1}\right)_{y+\upsilon,z+\kappa} \tag{10}$$

where, * denotes the convolutional operator for extracting the local patterns using the output of the alternate layers in conv layer, and $(V_u^j)_{y,z}$ symbolizes the fixed feature map. The feature maps from the previous conv layer f_1^{q-1} form the input to the j^{th} conv layer. The arbitrary filter $\left(\beta_{u,q}^j\right)_{\upsilon,\kappa}$ specifies the weights of j^{th} conv layer, which is trained using the proposed PI algorithm. The filter $\beta_{u,q}^j$ links the q^{th} feature map of $(j-1)^{th}$ conv layer and the u^{th} feature map in j^{th} conv layer. The bias matrix of the corresponding j^{th} conv layer is, b_u^j. ReLU is the activation function for eradicating the negative values toassist simplicity and effectiveness. The j^{th} non-linear layer is fed with the feature maps, and the output is given in (11).

$$V_u^j = fn\left(V_q^{j-1}\right) \tag{11}$$

where $fn(\,)$ denotes the activation function in convlayer j. The ReLU layer is significant because it allows Deep CNN to run faster and handle a large number of networks.

Pooling (POOL) layers

Complexity is reduced by the POOL layer, which includes neurons connecting the square-shaped space along its width and height of the preceding levels. The POOL layers perform a specified action even though they lack bias or training weights.

Fully connected layers: Abstract characteristics from the pooling and conv layers make up the input to the fully connected layer. The outcome of the fully connected layer can be seen in (12).

$$R_u^j = \delta(X_u^j) \text{ with } X_u^j = \sum_{q=1}^{f_1^{j-1}} \sum_{y=1}^{f_2^{j-1}} \sum_{z=1}^{f_3^{j-1}} \left(\varpi_{u,q,y,z}^j\right)\left(V_q^{j-1}\right)_{y,z} \tag{12}$$

where $\varpi_{u,q,y,z}^j$ refers to the weight that connects the unit (y, z) in the q^{th} feature map of the layer $(j-1)$ and u^{th} unit in the layer j.

b) Training Phase: Steps to Determine the Optimal Weights for Deep CNN

As shown in fig. 4, the algorithmic procedures for determining the best weights to train the deep CNN are explored further below.

a) Initialization: The swarm population is initialized in the first phase, and let's assume that there are a certain number of solutions, as shown in (13).

$$Z_a; (1 \le a \le N) \tag{13}$$

*b) Fitness Evaluation:*The fitness of the solutions is evaluated for individual iteration to choose the optimal solution. The fitness depends on the minimal error that is obtained by taking the square of the absolute difference of the estimated and the target output The MSE error is given in (14).

$$\varepsilon = \frac{1}{M}\sum_{e=1}^{M}|E - O|^2 \tag{14}$$

where E is the estimated output and O is classifier output. The total input samples are denoted as, M.

c) Computing the optimal position based on the minimum value of the error: The position of the particles using the minimal error. The position update follows either of the algorithms, namely the PI algorithm and Stochastic Gradient Descent (SGD) [36] algorithm. Whenever the error corresponding to the PI algorithm is less than the error corresponding to the SGD algorithm, the position of the particles is updated based on the PI algorithm, or else the position update follows the SGD algorithm. The position update is based on (15),

$$Z^{\tau+1} = \begin{cases} Z_{SGD}^{\tau+1} & ;i \quad f\varepsilon^{SGD} < \varepsilon^{PI} \\ Z_{PI}^{\tau+1} & ;O \quad therwise \end{cases} \tag{15}$$

where $Z_{SGD}^{\tau+1}$ refers to the position update based on SGD algorithm that is given in (16).

$$Z_{SGD}^{\tau+1} = \left(1 - \frac{1}{\tau}\right) Z^{\tau} + Y \tag{16}$$

where Y indicates the training sample and $Y \in (A, B)$. A specifies the input feature vector and B refers to the categories for the individual training sample. The position updated based on the PI algorithm is denoted as, $Z_{PI}^{\tau+1}$ and the derivation of the proposed equation follows as below:

i) Update equation for the PI algorithm: The optimal weights for training the deep CNNare derived using PI optimization. The position update of the particles is based on the PI algorithm and the update equation is obtained through modifying the PSO equation with the Imperialist Colony algorithm. The standard equation for the PSO algorithm is given in (17).

$$Z^{\tau+1} = Z^{\tau} + v^{\tau+1} \tag{17}$$

Substitute the velocity of the particle in (17), the position update is modified and it is given in (18).

$$Z^{\tau+1} = Z^{\tau} + \omega v^{\tau} + h_1 t_1 (P^{\tau} - Z^{\tau}) + h_2 t_2 (\Omega^{\tau} - Z^{\tau}) \tag{18}$$

where Z^{τ} refers to the position of the particle at a time τ. Ω^{τ} and P^{τ} are the global and the personal best solutions of the particles at iteration τ. h_1 and h_2 are the constants specifying the cognitive learning factor and acceleration factor, and t_1, t_2 are the random numbers. v^{τ} indicates the velocity of the particle at iteration τ and ω specifies the inertial weight. Equations (19), (20), and (21) are obtained by rearranging the equation (18), the position of the particle at iteration τ is derived based on the local and the global position of the particles as given in equation (21).

$$Z^{\tau+1} = Z^{\tau} + \omega v^{\tau} + h_1 t_1 P^{\tau} - h_1 t_1 Z^{\tau} + h_2 t_2 \Omega^{\tau} - h_2 t_2 Z^{\tau} \tag{19}$$

$$Z^{\tau+1} = Z^{\tau}[1 - h_1 t_1 - h_2 t_2] + \omega v^{\tau} + h_1 t_1 P^{\tau} + h_2 t_2 \Omega^{\tau} \tag{20}$$

$$Z^{\tau} = \frac{1}{[1 - h_1 t_1 - h_2 t_2]}\left[Z^{\tau+1} - \omega v^{\tau} - h_1 t_1 P^{\tau} - h_2 t_2 \Omega^{\tau}\right] \tag{21}$$

The standard equation of the Imperialist Colony algorithm is given as in equation (22) that depicts that the position at iteration $\tau + 1$ is based on the global best position and the position of the particle at τ.

$$Z^{\tau+1} = \Omega^{\tau} + 2\gamma\varphi Z^{\tau} \tag{22}$$

where γ denote the constant and its value exceeds 1 and ϕ denotes the distance among the colony and imperialist countries. Ω^{τ} signifies the global best solution and $Z^{\tau+1}$ is the position at the time $\tau+1$. Substituting equation (21) in equation (22), the equations (23) to (27) are obtained.

$$Z^{\tau+1} = \Omega^{\tau} + \frac{2\gamma\varphi}{[1-h_1t_1-h_2t_2]}\left[Z^{\tau+1} - \omega v^{\tau} - h_1t_1P^{\tau} - h_2t_2\Omega^{\tau}\right] \quad (23)$$

$$Z^{\tau+1} = \Omega^{\tau} + \frac{2\gamma\varphi\times Z^{\tau+1}}{[1-h_1t_1-h_2t_2]} - \frac{\omega v^{\tau}\times 2\gamma\varphi}{[1-h_1t_1-h_2t_2]} - \frac{h_1t_1P^{\tau}\times 2\gamma\varphi}{[1-h_1t_1-h_2t_2]} - \frac{h_2t_2\Omega^{\tau}\times 2\gamma\varphi}{[1-h_1t_1-h_2t_2]} \quad (24)$$

$$Z^{\tau+1} - \frac{2\gamma\varphi\times Z^{\tau+1}}{[1-h_1t_1-h_2t_2]} = \Omega^{\tau} - \frac{2\gamma\varphi}{[1-h_1t_1-h_2t_2]}\{\omega v^{\tau} + h_1t_1P^{\tau} + h_2t_2\Omega^{\tau}\} \quad (25)$$

$$Z^{\tau+1}\left[\frac{1-h_1t_1-h_2t_2-2\gamma\varphi}{1-h_1t_1-h_2t_2}\right] = \Omega^{\tau} - \frac{2\gamma\varphi}{[1-h_1t_1-h_2t_2]}\{\omega v^{\tau} + h_1t_1P^{\tau} + h_2t_2\Omega^{\tau}\} \quad (26)$$

$$Z^{\tau+1} = \frac{1-h_1t_1-h_2t_2}{1-h_1t_1-h_2t_2-2\gamma\varphi}\left\{\Omega^{\tau} - \frac{2\gamma\varphi}{[1-h_1t_1-h_2t_2]}\{\omega v^{\tau} + h_1t_1P^{\tau} + h_2t_2\Omega^{\tau}\}\right\} \quad (27)$$

The equation (27) makes it abundantly evident that the particle's position in the current iteration is dependent on its location, individual best, and overall best position in the previous iteration.

d) Stopping criterion: Updating positions is done repeatedly until the best overall solution is found.

Algorithm for deriving optimal weights to tune deep CNN

Optimal weight formulation for deep CNN

1 **Input**: Swarm Population Z_a; $(1 \le a \le N)$
2 **Output**: Best position of the particle $Z^{\tau+1}$
3 Swarm Initialization
4 Evaluation of the fitness ε
5 #Position Update
6 *If* $(\varepsilon^{SGD} < \varepsilon^{PI})$
7 {
8 Update the position of the particle using equation

$$Z_{SGD}^{\tau+1} = \left(1 - \frac{1}{\tau}\right) Z^{\tau} + Y$$

9 Else
10 Update the position based on equation

$$Z^{\tau+1} = \frac{1-h_1t_1-h_2t_2}{1-h_1t_1-h_2t_2-2\gamma\varphi}\left\{\Omega^{\tau} - \frac{2\gamma\varphi}{[1-h_1t_1-h_2t_2]}\{\omega v^{\tau} + h_1t_1P^{\tau} + h_2t_2\Omega^{\tau}\}\right\}$$

11 }
12 Repeat steps 4 to 11
13 End

IV. Results and Discussion

In order to demonstrate the efficiency of the suggested approach in estimating tumor levels, the section in question deliberates on the findings of the produced method when compared to the existing methods.

A. Setup Used

The PI-Deep CNN is implemented using MATLAB software operating on PC with Windows 8OS. The BRATS database, where the MRI images specific to each patient are maintained and there are four different modalities, including T1, T2, T1C, and FLAIR for each patient, is used to test and assess the approaches [37]. It also comprises feature and textural patterns, such as co-occurrence matrices, specified block sizes, mean and variance of slice or radial distance, etc. Performance is considerably improved. For every segmentation task, fixed groupings of algorithmic segmentations consistently outperformed the best individual segmentation algorithm. Online evaluation serves as the BRATS benchmark's main component.

B. Segmentation Output

The sample results are demonstrated in Fig.3 and Fig. 4, respectively that show the results obtained using the SLNS descriptor. Fig.3depicts the segmentation output of the SLNS descriptor using image 1. Slice 76 and 106 of the input image 1are depicted in fig.3 a) and fig.3 c), respectively. The segmented output for slice 76 and slice 106 are demonstrated in fig.3 b) and fig.3 d), respectively.

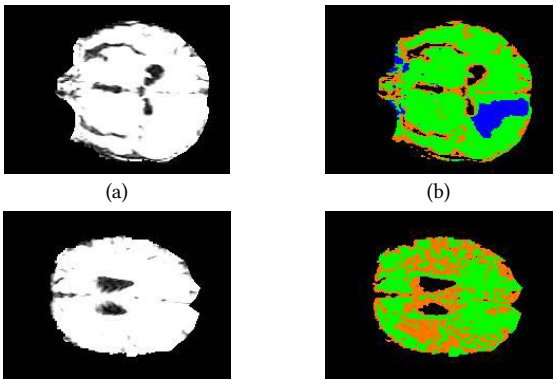


Fig. 3. Segmentation results using Input image 1 (a) Slice 76 of Input image 1, (b) Segmented output of (a), (c) Slice 106 of Input image 1, d) Segmented output of (c).

Fig.4 shows the segmentation results using image 2. The slices 116 and 81 of image 2are pictured in Fig.4 a) and Fig.4 c), respectively, whereas the segmented output of the slices is pictured in Fig.4b) and Fig. 4 d), respectively.

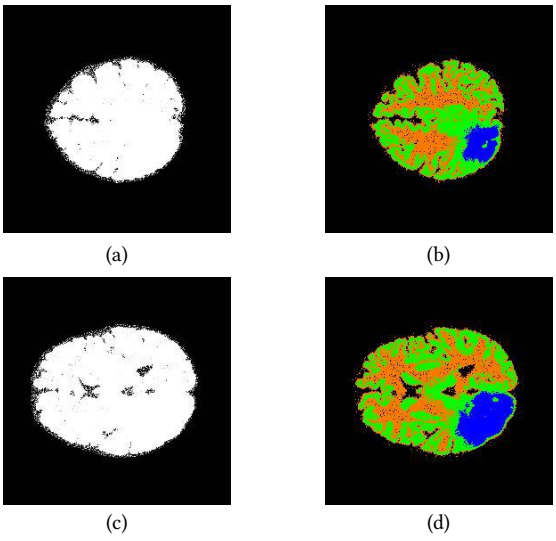
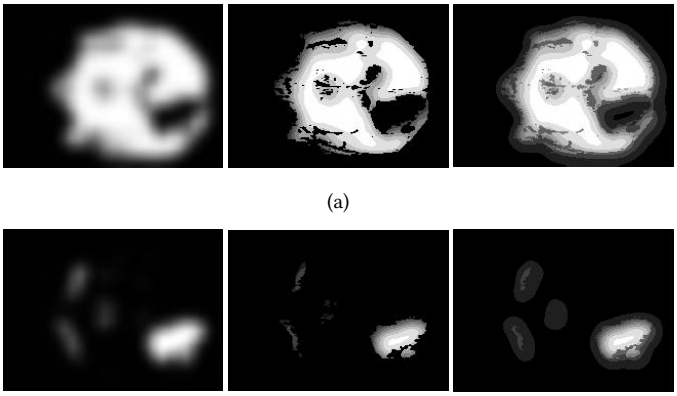


Fig. 4. Segmentation results using Input image 2 a) Slice 116 of Input image 2, b) Segmented output of (a), c) Slice 81 of Input image 2, d) Segmented output of (c).

The tumor and non-tumor zones that the proposed SLNS descriptor identified are shown in FIGS. 5A and 5B, respectively. To guarantee the classification accuracy of the classifiers, the input MRI image's tumor and non-tumor regions are subjected to feature extraction.



(a)

(b)

Fig. 5.Results of proposed SLNS descriptor a) Tumor region b) Non-tumor region.

C. Competing Methods

In order to demonstrate the effectiveness of the proposed method, it is compared with a number of other methods, including Convolutional Neural Networks (CNN) [6], PSSW+MCSS [16], Principle Component Analysis (PCA) [9], GLCM-CA [10], Markov-random field [15], Deep Neural Networks (DBN) [25], KNN [38], and NN [39].

D. Comparative Analysis

Sensitivity, specificity, accuracy, and ROC are used in a comparative examination of the suggested classifier's performance using the BRATS database. The suggested classifier is contrasted with techniques like KNN, NN, PCA, DBN, CNN, and PI-deep CNN in order to evaluate the performance effectively.

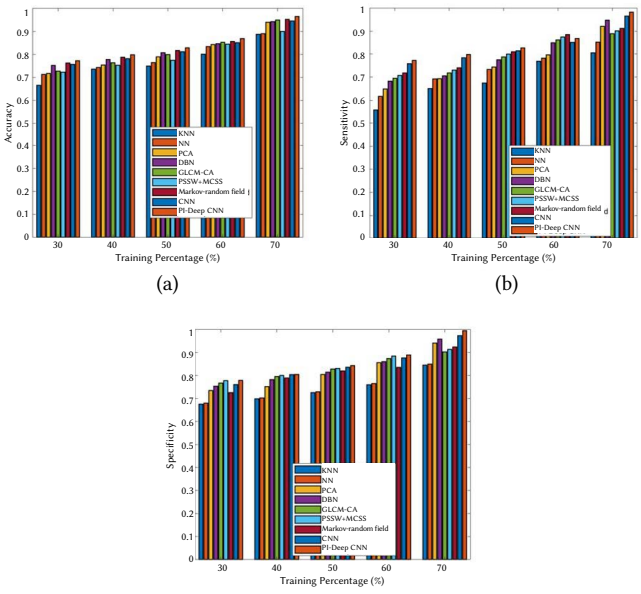
a) Analysis Using Image 1

Figures 6.a, 6.c, and 6.b, respectively, show the analysis of accuracy, specificity, and sensitivity using image 1 for the training percentage. For a training percentage of 70%, the accuracy of the comparative methods KNN, NN, PCA, DBN, GLCM-CA, PSSW+MCSS, Markov-random field, CNN, and PI-Deep CNN is 0.8871%, 0.89%, 0.94%, 0.942%, 0.95%, 0.899%, 0.952%, 0.945%, and 0.965%, respectively. This demonstrates that the proposed method is more accurate than all of the methods. The suggested technique is superior to the existing methods, as shown by the sensitivity of KNN, NN, PCA, DBN, GLCM-CA, PSSW+MCSS, Markov-random field, CNN, and PI-Deep CNN, which is 0.8046%, 0.8533%, 0.9214%, 0.9470%, 0.8896%, 0.9019%, 0.9122%, 0.965%, and 0.9821%, respectively, for 70% training data. For 70% training data, the specificity of the competing methods—KNN, NN, PCA, DBN, GLCM-CA, PSSW+MCSS, Markov-random field, CNN, and PI-Deep CNN—is, respectively, 0.8448%, 0.85%, 0.9413%, 0.958%, 0.9025%, 0.9131%, 0.9245%, 0.9782%, and 0.9951%.

Figures 7.a, 7.b, and 7.c show the analysis using picture 1, respectively. According to the cross-fold validation 10 results, the accuracy of the comparative methods KNN, NN, PCA, DBN, GLCM-CA, PSSW+MCSS, Markov-random field, CNN, and PI-Deep CNN is 0.8845, 0.8873, 0.9371, 0.9391, 0.9471, 0.8945, 0.9491, 0.9421, and 0.9621, respectively. This demonstrates that the proposed method is more accurate than the existing methods. The cross-fold validation10 shows that the proposed method outperforms the existing methods in terms of sensitivity with sensitivity values of 0.7749, 0.8465, 0.9140, 0.9394, 0.8923, 0.9046, 0.9149, 0.9572, and 0.9742, respectively, for the comparative methods of KNN, NN, PCA, DBN, GLCM-CA, PSSW+MCSS, Markov-random field, CNN, and PI-Deep. For the cross-fold validation 10, the specificities of KNN, NN, PCA, DBN, GLCM-CA, PSSW+MCSS, Markov-random field, CNN, and PI-Deep CNN are, respectively, 0.8397, 0.85, 0.9356, 0.9522, 0.9125, 0.9261, 0.9397, 0.9679, and 0.9892. Thus, it has been established that PI-Deep CNN performs better in terms of sensitivity, specificity, and accuracy than the current approaches.

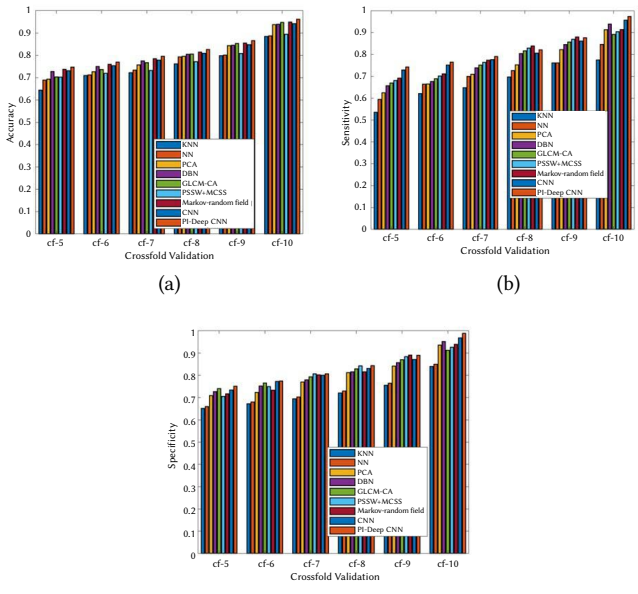
b) Analysis Using Image 2

The analysis based on the metrics, like accuracy, specificity, and sensitivity using image 2 based on the training percentage is depicted in Fig.8.a, Fig. 8.c, and Fig. 8.b,respectively. The accuracy of the comparative methods, KNN, NN, PCA, DBN, GLCM-CA, PSSW+MCSS, Markov-random field, CNN, and PI-Deep CNN, is 0.7670%, 0.89%, 0.94%, 0.9424%, 0.9374%, 0.9355%, 0.9305%, 0.945%, and 0.9469%, respectively, for 70% training data, which proves that the proposed method is superior over the existing methods in terms of accuracy.



(a) (b) (c)

Fig. 6. Comparative analysis based on training percentage using the image 1 a) accuracy b) sensitivity c) specificity



(a) (b) (c)

Fig. 7. Comparative analysis based on cross-fold using the image 1 a) accuracy b) sensitivity c) specificity.

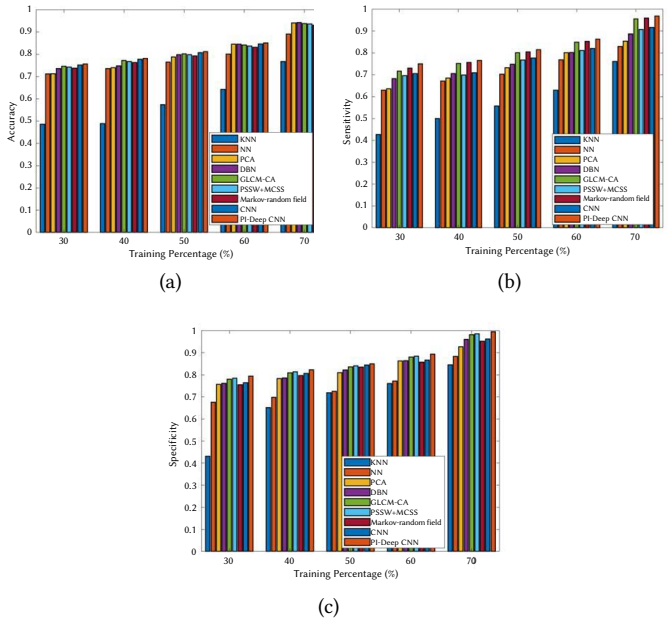


Fig. 8. Comparative analysis based on training percentage using the image 2 a) accuracy b) sensitivity c) specificity.

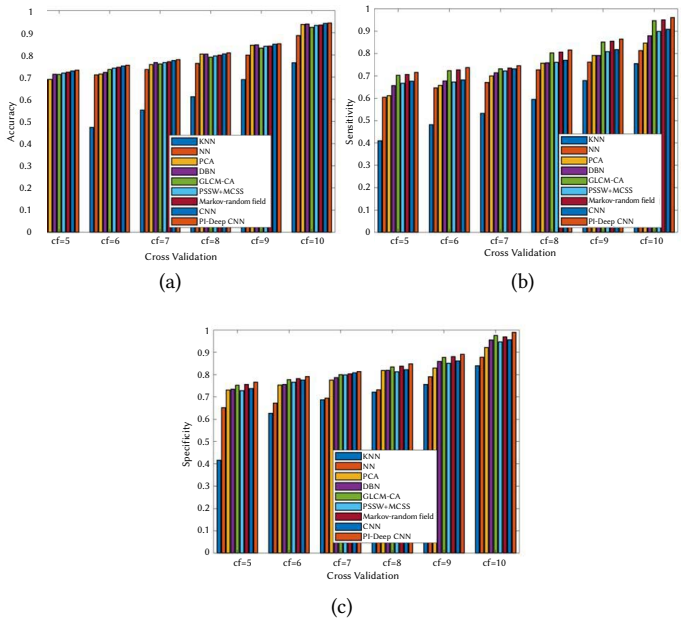


Fig. 9. Comparative analysis based on cross-fold using the image 2 a) accuracy b) sensitivity c) specificity.

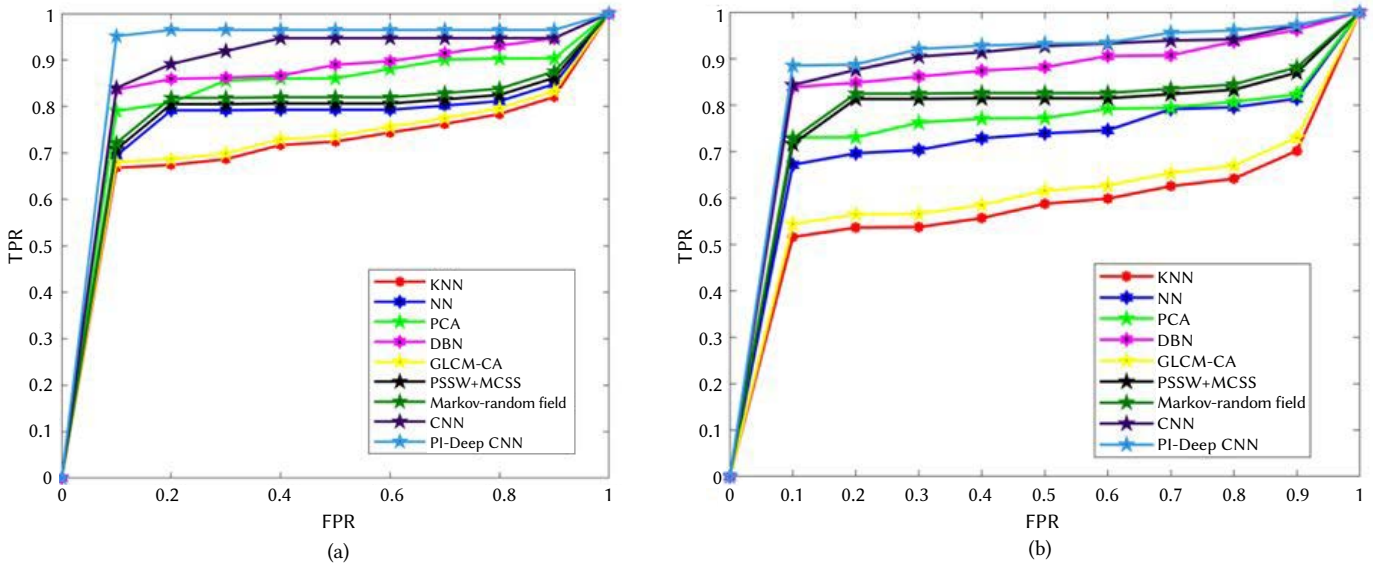


Fig. 10. ROC a) Image 1 b) Image2.

The sensitivity of the comparative methods, KNN, NN, PCA, DBN, GLCM-CA, PSSW+MCSS, Markov-random field, CNN, and PI-Deep CNN, is 0.7612%, 0.8291%, 0.8533%, 0.8866%, 0.9549%, 0.9066%, 0.9589%, 0.9161%, and 0.9684%, respectively, for 70% training data, which proves that the proposed method is superior over the existing methods in terms of sensitivity. The specificity of the comparative methods, KNN, NN, PCA, DBN, GLCM-CA, PSSW+MCSS, Markov-random field, CNN, and PI-Deep CNN, is 0.8448%, 0.8831%, 0.9271%, 0.9604%, 0.9816%, 0.9856%, 0.9523%, 0.9618%, and 0.9951%, respectively, for 70% training data, which proves that the proposed method is superior over the existing methods in terms of sensitivity.

Figures 9.a, 9.b, and 9.c, respectively, display the analysis of accuracy, sensitivity, and specificity utilizing image 2 based on the cross-fold validation. According to the cross-fold validation 10 results, the accuracy of the comparative methods KNN, NN, PCA, DBN, GLCM-CA, PSSW+MCSS, Markov-random field, CNN, and PI-Deep CNN is 0.7647, 0.8873, 0.9371, 0.9395, 0.9246, 0.9326, 0.9346, 0.9421, and 0.9441, respectively. This demonstrates that the proposed method is more accurate than the existing methods. The cross-fold validation 10 shows that the proposed method outperforms the existing methods in terms of sensitivity with sensitivity values of 0.7551, 0.8135, 0.8465, 0.8795, 0.9472, 0.8993, 0.9512, 0.9088, and 0.9607, respectively, for the comparative methods of KNN, NN, PCA, DBN, GLCM-CA, PSSW+MCSS, Markov-random field, CNN, and PI-Deep The KNN, NN, PCA, DBN, GLCM-CA, PSSW+MCSS, Markov-random field, CNN, and PI-Deep CNN comparative methods have specificities of 0.8397, 0.8778, 0.9216, 0.9546, 0.9757, 0.9466, 0.9685, 0.9561, and 0.9892, respectively, for the cross-fold validation 10; this demonstrates that the proposed method is more sensitive than the existing methods.

c) *Analysis of ROC*

The ROC curve is shown using images 1 and 2, respectively, in Figures 10.a and 10.b. The TPR for the minimal value 0.1 of FPR is 0.9514, as can be observed in Fig. 10.a, but for methods like KNN,

NN, PCA, DBN, GLCM-CA, PSSW+MCSS, Markov-random field, and CNN, it is 0.6680, 0.6945, 0.7905, 0.8358, 0.6805, 0.7080, 0.7214, and 0.84. Similar to this, in fig.10.b, the TPR for the proposed approach is 0.8853 when the FPR of the PI-Deep CNN is at a minimum value of 0.1, but it is 0.5158, 0.6716, 0.7302, 0.8388, 0.5441, 0.7162, 0.7277, 0.8435, and 0.8435 for methods like KNN, NN, PCA, DBN, GLCM-CA, PSSW+MCSS, Markov-random field, and CNN.

E. Analysis Using the Best Performance of the Comparative Methods

The analysis employing the tumor categorization techniques based on performance indicators is shown in Table I. For 70% training data, the accuracy of the methods KNN, NN, PCA, DBN, GLCM-CA, PSSW+MCSS, Markov-random field, CNN, and suggested PI-Deep CNN, respectively, is 0.8871%, 0.89%, 0.945, 0.942%, 0.95%, 0.899%, 0.952%, 0.945%, and 0.965%. The accuracy of PI-Deep CNN is 0.9621 for the cross-fold 10, compared to 0.8845, 0.8873, 0.9371, 0.9391, 0.9471, 0.8945, 0.9491, 0.9421, and 0.9621 for KNN, NN, PCA, DBN, GLCM-CA, PSSW+MCSS, Markov-random field, and CNN, respectively. Table 1 makes it evident that the proposed method, which was based on cross-fold validation and the training percentage, obtained the highest levels of accuracy, sensitivity, and specificity.

TABLE I. Discussion of the Comparative Methods

		Training percentage =70		
		Sensitivity	**Accuracy**	**Specificity**
Based on training percentage	KNN	0.8046	0.8871	0.8448
	NN	0.8533	0.89	0.85
	PCA	0.9214	0.94	0.9413
	DBN	0.947	0.942	0.958
	GLCM-CA	0.889	0.95	0.9025
	PSSW+MCSS	0.901	0.899	0.9131
	Markov-random field	0.9122	0.952	0.9245
	CNN	0.965	0.945	0.9738
	PI-Deep CNN	0.9821	0.965	0.9951
Based on the Crossfold		**Crossfold validation =10**		
	Methods	**Sensitivity**	**Accuracy**	**Specificity**
	KNN	0.7749	0.8845	0.8397
	NN	0.8465	0.8873	0.85
	PCA	0.9140	0.9371	0.9356
	DBN	0.9394	0.9391	0.9522
	GLCM-CA	0.8923	0.9471	0.9125
	PSSW+MCSS	0.9046	0.8945	0.9261
	Markov-random field	0.9149	0.9491	0.9397
	CNN	0.9572	0.9421	0.9679
	PI-Deep CNN	0.9742	0.9621	0.9892

V. Conclusion

Using a novel brain tumor segmentation framework and MRI brain tumor images, the segmentation and classification of brain tumors is advanced. The main step that is necessary for the accurate classification of the level of brain tumor is segmentation. The DBSCAN Clustering Algorithm is used to partition the brain picture, creating clusters that are then subjected to feature extraction. The LNS descriptor and scattering transform are combined in SLNS, which captures textural characteristics for accurate tumor level categorization. The grid-based shape features, which are retrieved from both the tumor and non-tumor regions, make up the feature vector. The Deep Convolutional Neural Network classifier, which was trained using the suggested PI technique, is given the features. In order for the doctor to make an accurate diagnosis, the tumor class is finally determined using the classifier as Normal without tumor, Abnormal, Malignant tumor, and Non-malignant tumor. The proposed method obtained an accuracy of 0.965, a sensitivity of 0.9821, and a specificity of 0.9951, according to the experimentation utilizing the BRATS database.

References

[1] Lubna Farhi, Adeel Yusuf, and Rana Hammad Raza, "Adaptive stochastic segmentation via energy-convergence for brain tumor in MR images," Journal of Visual Communication and Image Representation, vol. 46, pp. 303-311, 2017.

[2] S. Bauer et al., "A survey of MRI-based medical image analysis for brain tumor studies," Phys. Med. Biol., vol. 58, no. 13, pp. 97–129, 2013.

[3] D. N. Louis et al., "The 2007 who classification of tumors of the central nervous system," Acta Neuropathologica, vol. 114, no. 2, pp. 97–109, 2007.

[4]]E. G. Van Meir et al., "Exciting new advances in neuro-oncology: The avenue to a cure for malignant glioma," CA, Cancer J. Clinicians, vol. 60, no. 3, pp. 166–193, 2010.

[5] G. Tabatabai et al., "Molecular diagnostics of gliomas: The clinical perspective," Acta Neuropathologica, vol. 120, no. 5, pp. 585–592, 2010.

[6] Sergio Pereira, Adriano Pinto, Victor Alves, and Carlos A. Silva, "Brain tumor segmentation using convolutional neural networks in MRI images," IEEE transactions on medical imaging, vol. 35, no. 5, pp. 1240-1251, 2016.

[7] P. John, et al., "Brain tumor classification using wavelet and texture based neural network," International Journal of Scientific & Engineering Research, vol. 3, no.l0, 2012.

[8] S. Charfi, R. Lahmyed, and Rangarajan, "A novel approach for brain tumor detection using neural network," International Journal of Research in Engineering and Technology, vol. 2, pp. 93-104.

[9] Irem Ersoz Kaya, Ayca Cakmak Pehlivanlı, Emine Gezmez Sekizkardes, and Turgay Ibrikci, "PCA based clustering for brain tumor segmentation of T1w MRI images," Computer Methods and Programs in Biomedicine, vol. 140, pp. 19-28, 2017.

[10] Chaiyanan Sompong, and Sartra Wongthanavasu, "An efficient brain tumor segmentation based on cellular automata and improved tumor-cut algorithm," Expert Systems with Applications, vol. 72, pp. 231-244, 2017.

[11] Cristina Utrilla Contreras, Nelson Mauricio, Buitrago Sánchez, Joachim Graessner, Pilar García Raya, and Begoña Marin Aguilera, "Difusion-Weighted MRI: from Brownian Motion to Head&Neck Tumor Characterization," International Journal of Interactive Multimedia and Artificial Intelligence, vol.4, no.5, pp.6-14, 2017.

[12] S. Bauer, R. Weiest, L. P. Nolte, and M. Reyes, "A survey of MRI based medical image analysis for brain tumor studies," Phy.Med. Biology, vol. 58, no. 13, pp. 97–129, 2013.

[13] Afsaneh Abdollahzadeh Rezaie and Ali Habiboghli, "Detection of Lung Nodules on Medical Images by the Use of Fractal Segmentation," International Journal of Interactive Multimedia and Artificial Intelligence, vol.4, no.5, pp.15-19, 2017.

[14] S. K. Warfield, K. H. Zou, and W. M. Wells, "Brain tumor segmentation from multispectral MRIs using sparse representation classification and Markov Random Field regularization," IEEE Trans. Med. Imaging, vol. 23, no. 7, pp. 903–921, 2004.

[15] Yuhong Li, Fucang Jia, and Jing Qin, "Brain tumor segmentation from multimodal magnetic resonance images via sparse representation," Artificial intelligence in medicine, vol. 73, pp. 1-13, 2016.

[16] Elisee Ilunga-Mbuyamba, Jorge Mario Cruz-Duarte, Juan Gabriel Avina-Cervantes, Carlos Rodrigo Correa-Cely, Dirk Lindner, and Claire Chalopin, "Active contours driven by Cuckoo Search strategy for brain tumor images segmentation," Expert Systems with Applications, vol. 56, pp. 59-68, 2016.

[17] Avinash Gopal, "Hybrid classifier: Brain Tumor Classification and Segmentation using Genetic-based Grey Wolf optimization," Multimedia Research, vol 3, No 2, 2020.

[18] Pierfrancesco Fusco, Vincenza Cofini, Emiliano Petrucci, Paolo Scimia, Giuseppe Paladini, Astrid U Behr, Fabio Gobbi, Tullio Pozone, Giorgio Danelli, Mauro Di Marco, Roberto Vicentini, Stefano Necozione, and

Franco Marinangeli, "Unilateral paravertebral block compared with subarachnoid anesthesia for the management of postoperative pain syndrome after inguinal herniorrhaphy a randomized controlled clinical trial," Pain, vol.157, no.5, pp.1105, 1113, 2016.

[19] MD Pierfrancesco Fusco, MD Vincenza Cofini, MD Emiliano Petrucci, MD Paolo Scimia, MD Tullio Pozone, MD Giuseppe Paladini, MD Gaspare Carta, and MD Stefano Necozione, "Transversus abdominis plane block in the management of acute postoperative pain syndrome after caesarean section: a randomized controlled clinical trial," Pain Physician, vol.19, pp.583-591, 2016.

[20] Nitin Deotale,Uttam Kolekar,Anuradha Kondelwar, "Self-adaptive Particle Swarm Optimization for Optimal Transmit Antenna Selection," Journal of Networking and Communication Systems, vol.3, no.1, pp.1-10,2020.

[21] Rupam Gupta Roy,Dibyendu Ghoshal, "Search and Rescue Optimization Algorithm - Second Order Sliding Mode Control: AUV Error Tracking," Journal of Computational Mechanics, Power System and Control, vol.3, no.1, pp.10-20,2020.

[22] T. Wang, I. Cheng, and A. Basu, "Fluid vector flow and application in brain tumor segmentation," IEEE Trans. Biomed. Eng., vol. 56, no .3, pp. 781–789, 2009.

[23] J. Sachdeva, V. Kumar, I. Gupta, N. Khandelwal, and C. K. Ahuja, "A novel content-based active model for brain tumor segmentation," Magn. Reson. Imaging, vol. 30, pp. 694–715, 2012.

[24] A. Hamamci, N. Kucuk, K. Karaman, K. Engin, and G. Unal, "Tumor-cut: segmenation of brain tumors on contrast enhanced MR images for radiosurgery applications," IEEE Trans. Med. Imaging, vol. 31, no. 3, pp. 790–804, 2012.

[25] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle, "Brain tumor segmentation with deep neural networks," Medical image analysis, vol. 35, pp. 18-31, 2017.

[26] F. Heydarpour, E. Abbasi, M. J. Ebadi, and S. M. Karbassi, "Solving an Optimal Control Problem of Cancer Treatment by Artificial Neural Networks," International Journal of Interactive Multimedia and Artificial Intelligence, vol.6, no.4, pp.18-25, 2020.

[27] Elisee Ilunga-Mbuyamba, Juan Gabriel Avina–Cervantes, Arturo Garcia–Perez, Rene de Jesus Romero–Troncoso, Hugo Aguirre–Ramos, Ivan Cruz–Aceves, and Claire Chalopin, "Localized active contour model with background intensity compensation applied on automatic MR brain tumor segmentation," Neuro computing, vol. 220, pp. 84-97, 2017.

[28] J. Shen, X. Hao, Z. Liang, Y. Liu, W. Wang, and L. Shao, "Real-Time Superpixel Segmentation by DBSCAN Clustering Algorithm," in IEEE Transactions on Image Processing, vol. 25, no. 12, pp. 5933-5942, Dec. 2016.

[29] Prateekshit Pandey, Richa Singh, and Mayank Vatsa, " Face recognition using scattering wavelet under Illicit Drug Abuse variations," In International Conference on Biometrics (ICB), pp.1-6, 2016.

[30] J. Shen, X. Hao, Z. Liang, Y. Liu, W. Wang, and L. Shao, "Real-Time Superpixel Segmentation by DBSCAN Clustering Algorithm," in IEEE Transactions on Image Processing, vol. 25, no. 12, pp. 5933-5942, Dec. 2016.

[31] Prateekshit Pandey; Richa Singh; and Mayank Vatsa, "Face Recognition using ScatteringWavelet under Illicit Drug Abuse Variations," International Conference on Biometrics (ICB), 25 August 2016.

[32] Dian Palupi Rini, Siti Mariyam Shamsuddin, and Siti Sophiyati Yuhaniz, "Particle Swarm Optimization: Technique, System and Challenges," International Journal of Computer Applications, vol. 14, no. 1, pp. 0975 – 8887, 2011.

[33] E. Atashpaz-Gargari and C. Lucas, "Imperialist competitive algorithm: An algorithm for optimization inspired by imperialistic competition," in Proceedings of IEEE Congress on Evolutionary Computation, pp. 4661-4667, 2007.

[34] Giduthuri Sateesh Babu, Peilin Zhao, and Xiao-Li Li, " Deep Convolutional Neural Network Based Regression Approach for Estimation of Remaining Useful Life," International Conference on Database Systems for Advanced Applications (DASFAA), pp. 214-228, 2016.

[35] Dong, Chao, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Image super-resolution using deep convolutional networks," IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 2, pp.295-307, 2016.

[36] Jing Yang and Guanci Yang, "Modified Convolutional Neural Network Based on Dropout and the Stochastic Gradient Descent Optimizer," vol.11, no.3, 2018.

[37] Bjoern H. Menze et al., "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," IEEE Transactions on Medical Imaging, Vol. PP, no. 99, pp.1, 2014.

[38] V.Vani and M. Kalaiselvi Geetha, "Automatic Tumor Classification of Brain MRI Images," International Journal of Computer Sciences and Engineering, vol.4, no.10, 2016.

[39] N. V. S. Natteshan and J. Angel Arul Jothi, "Automatic Classification of Brain MRI Images Using SVM and Neural Network Classifiers," Advances in Intelligent Informatics, vol.320, pp.19-30, 2015.

Maahi Amit Khemchandani

Maahi Amit Khemchandani has completed Bachelor's Degree from North Maharashtra University, Master's Degree from Bharati Vidyapeeth College of Engineering, Mumbai University, Mumbai and Research Scholar at Dr. Babasaheb Ambedkar Technological University, Lonere . She is having more than 12 years of experience in teaching field. Her area of interest is in Networking and Machine Learning.

Shivajirao Manikrao Jadhav

Dr. Shivajirao Manikrao Jadhav has completed his P.hd from Dr. B.A.T.U, Lonere. He is now working as Professor and Head of Department of Information Technology at Dr. Babasaheb Ambedkar Technological University, Lonere. He is having teaching experience over 20 years and providing his valuable guidance to students in many research areas like Soft Computing, Machine Learning, Hybrid Neural Networks and Biomedical Signal Analysis. He has received UGC's Teacher Fellowship award under FIP. He has conducted many workshops. He is Chairman of Board of Studies at Dr. B.A.T.U, Lonere.

B. R. Iyer

Dr. B. R. Iyer has completed his P.hd from Indian Institute of Technology, Roorkee. He is now working as Assistant Professor in Electronics & Telecommunication Engineering Department at Dr. Babasaheb Ambedkar Technological University, Lonere. He is providing his valuable guidance to students in many research areas like NANO Photonics/RF Front End Design Pervasive Healthcare System Design and allied Signal/ Image Processing, Wireless Sensor Network Design / IoT Education & Technology. He has received Indian National Academy of Engineers (INAE) research fellowship-2015-2016. He has honoured with Best research paper award at IEEE MTTs IMaRC-2013 held in New Delhi during 14-16 December 2013. He is Reviewer of IETE Journal of Research, Int. Jr. of RF & Microwave CAD(Willey), IJSAEM (Springer). He is Member, Technical Program Committee of APACE-14, ICACCI-14, ICCME-15, MALSPI-15, CNTIA-15, ICTEC-15, MOBIAPPS-15, I4CT-15, ICCCA-15.

Optimizing Fast Fourier Transform (FFT) Image Compression using Intelligent Water Drop (IWD) Algorithm

Surinder Kaur[1,2], Gopal Chaudhary[2]*, Javalkar Dinesh Kumar[1], Manu S. Pillai[2], Yash Gupta[2], Manju Khari[3]*, Vicente García-Díaz[4], Javier Parra Fuente[5]

[1] Lingaya's Vidyapeeth, Haryana (India)
[2] Bharati Vidyapeeth's College of Engineering, New Delhi (India)
[3]School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi (India)
[4] Departament of Computer Science, University of Oviedo (Spain)
[5] School of Engineering and Technology, Universidad Inernacional de La Rioja (UNIR), Logroño (Spain)

Received 2 February 2021 | Accepted 3 November 2021 | Published 21 January 2022

Abstract

Digital image compression is the technique in digital image processing where special attention is provided in decreasing the number of bits required to represent a digital image. A wide range of techniques have been developed over the years, and novel approaches continue to emerge. This paper proposes a new technique for optimizing image compression using Fast Fourier Transform (FFT) and Intelligent Water Drop (IWD) algorithm. IWD-based FFT Compression is a emerging methodology, and we expect compression findings to be much better than the methods currently being applied in the domain. This work aims to enhance the degree of compression of the image while maintaining the features that contribute most. It optimizes the FFT threshold values using swarm-based optimization technique (IWD) and compares the results in terms of Structural Similarity Index Measure (SSIM). The criterion of structural similarity of image quality is based on the premise that the human visual system is highly adapted to obtain structural information from the scene, so a measure of structural similarity provides a reasonable estimate of the perceived image quality.

Keywords

Structural Similarity Index, Fast Fourier Transform, Intelligent Water Drop, Image Compression.

DOI: 10.9781/ijimai.2022.01.004

I. Introduction

The application of images is expanding very rapidly. It is indispensable in the fields of remote sensing [1], video processing [2], medical science [3], machine/robot vision [4], and many other applications. Considering the rapid growth in the application scope of images, there is an increased demand for mass information storage and fast communication links [5].

Image compression applications use multiple methods and algorithms for compressing images, such as JPEG2000 [6], EBCOT [10], 2-D wavelet transformation [11]. The methods thus used can be categorized as lossless and lossy compression for image compression applications.

The compression method employed depends on the quality of the necessary output. If the application for image compression is expected to produce a very high-quality output without any loss of fidelity, then the lossless compression method is used. The lossy compression [12] method is used in applications where some quality can be compromised. There is a slight loss of quality in lossy compression, but the loss is too small to be noticed in terms of structural resemblance (SSIM) index and visual resemblance [19]–[21].

The digital image processing [7]–[9] techniques such as image sharpening and restoration, transmission and encoding, pattern recognition are optimized by maximizing the compression rate while maintaining an optimum percentage of data required to reconstruct the image with the highest quality [13]. The proposed optimized approach ensures quality along with efficient memory utilization. The proposed system aims to produce a compact image representation by reducing the requirements for image storage transmission, and processing. Malathkar et.al proposed an image compression algorithm consisting of a new simplified YUV colour space, corner clipping, uniform quantization, subsampling, differential pulse code modulation and Golomb Rice code for wireless capsule endoscopy [22]. In [23], the authors evaluate various compression models on their complexity and efficiency using various E-Learning images (Colour and Grey Scale) with different compression quality measurements.

Traditional image compression methods compress an image to such an extent that the decompressed image has much less structural similarity (SSIM) index [14] to the original one. The

* Corresponding author.
E-mail address: gopal.chaudhary88@gmail.com (G. Chaudhary), manjukhari@yahoo.co.in (M. Khari).

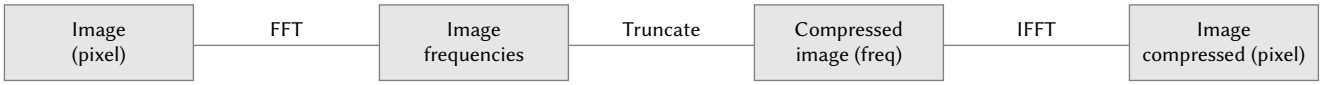


Fig. 1. Block diagram for image compression using Fast Fourier Transform.

procedure as visualised in Fig. 1, typically starts with applying Fast Fourier Transform(FFT) to the image and truncating the frequency domain output with specified threshold (independant of image content) followed by applying an inverse FFT to generate the compressed image. The proposed system uses fast fourier transform (FFT) [15] with optimized threshold values for each colour channel to compress the image that the decompressed image obtains high structural similarity (SSIM) index, thus extracting the contributing characteristics of the image.

A. Motivation

This section addresses the two factors that motivate the research undertaken in this study. First, there is plenty of research on optimizing compression techniques, but none have discussed concurrently compressing an image to its maximum level while sustaining its best possible quality. Second, there is a need to enhance the storage and transmission of compressed images. At times the storage and transmission costs are very high, to reduce the cost and fulfil the requirements of targeted process combination of optimum compression rate and quality must be achieved.

This paper is organized as follows. Section II explains the intelligent water drop (IWD) algorithm. Section III describes the image compression using fast Fourier transform (FFT). Section IV explains about structural similarity index (SSIM). Section V shows the proposed system. Section VI explains experimental results obtained using several sample images. Section VII concludes the complete research.

II. Intelligent Water Drop Algorithm

The algorithm Intelligent Water Drop (IWD) is motivated by studying the actual behaviour of natural drops in a flowing water source from elevated altitude to low altitude areas. A massive collection of drops governs water flow, each moving based on a shorter and simpler path naturally influenced, although subject to several environmental constraints. Shah-Hosseini expanded this basic idea to introduce the Intelligent Water Drop (IWD) algorithm for Traveling Salesman Problem (TSP) [16].

An IWD consists of two significant properties, similar to a natural water drop. These are a) the $IWD-soil(IWD)$ soil content and b) the $IWD-vel(IWD)$ velocity. The IWD's soil and velocity content dynamically change depending on the same route as it flows through the problem's discrete landscape. Therefore, depending on the movement of the IWD, some soil is removed from the traversed path and the corresponding soil path is dynamically updated in the process. Such flow leads to soil content decrease in ideal paths depending on the problem's setting. Thus it can be said that the routes with reduced soil content may be the most relevant to finding an almost ideal solution. Thus, the building of an ideal solution to the issue is governed by a set of evolving swarm behaviour linked to IWDs.

Concerning the original formulation of TSP problems, we can consider a graph $G = (V, E)$ where V is the set of nodes and E is the set of edges. Thus an IWD can be randomly positioned at any node. Say i, it follows the transition of probability as given in equation (1) to select the next node j.

$$P(i,j) = \frac{f(soil(i,j))}{\sum_{k \notin vc(IWD)} f(soil(i,k))} \tag{1}$$

$$f(soil(i,j)) = \frac{1}{\epsilon + g(soil(i,j))} \tag{2}$$

$$g(soil(i,j)) = soil(i,j) \;\; if \;\; minsoil \geq 0 \tag{3}$$

$$soil(i,j) - minsoil \quad if \quad minsoil < 0 \tag{4}$$

P(i, j) shows the transition probability of node j. K denotes exactly all nodes to be visited and π is the parameter of the algorithm. Thus, a node selection depends on the quantity of soil present on the edges among adjacent nodes given by soil(i,j) in a probabilistic way. Here min soil shows the least amount of soil on a path between any node i and j. The state transition probability of an IWD, as illustrated in equations (1) to (3), is therefore proportional to the soil content available in the edge between nodes i and j. As a result, as a path's soil content decreases, the probability of selecting the appropriate component of the solution increases. While each IWD moves incrementally from one node I to j while building a solution, the soil content of the IWD (soil(iwd)) and the velocity of the same (level(iwd)) is also updated based on equations (5-7).

$$\Delta vel^{IWD}(t) = \frac{a_v}{b_v + c_v \times soil^{2\alpha}(i,j)} \tag{5}$$

$$\Delta soil(i,j) = \frac{a_s}{b_s + c_s \times times^{2\beta}(i,j)} \tag{6}$$

$$time(i,j) = \frac{HUD(i,j)}{vel(IWD)} \tag{7}$$

HUD is a heuristic that can be used to measure an IWD's desirability/ unwantedness to select an edge between *i* and *j*, in this case. Therefore, a higher IWD velocity helps minimize the time an IWD takes to move from *i* to *j*. In turn, the time factor influences the amount of soil from a path to be removed (as shown in equation 5). The soil content of the entire solution path can be updated based on Equation (8) once the IWD attributes are calculated.

$$soil(i,j) = \rho_0 \times soil(i,j) - \rho_n \times \Delta soil(i,j) \tag{8}$$

Where ρ_0 and ρ_n remain within 0 and 1, according to the original TSP IWD algorithm, $\rho_0 = 1 - \rho_n$.

III. Fast Fourier Transform (FFT) for Image Compression

The Fourier transformation (FT) decomposes a time (a signal) into its constituent frequencies (also called analysis). This is similar to how a musical can be expressed in terms of its constituent notes volumes and frequencies (or pitches). The term Fourier transform refers to a function of time, both the representation and the mathematical operation associating the representation of the frequency domain.

A fast Fourier transformation (FFT) is an algorithm calculating the discrete Fourier transformation (DFT) or its inverse(IDFT) of a sequence. Analysis of Fourier converts a signal from its original domain in the frequency domain (often time or space) and vice versa. The DFT is obtained by breaking down a sequence of values into different components of the frequency.

Consider the pixel space image and apply a Fast Fourier Transform (FFT) to get a frequency domain image . For RGB layers, threshold values are calculated by truncating values below the calculated threshold, resulting in a compressed image in the frequency domain.

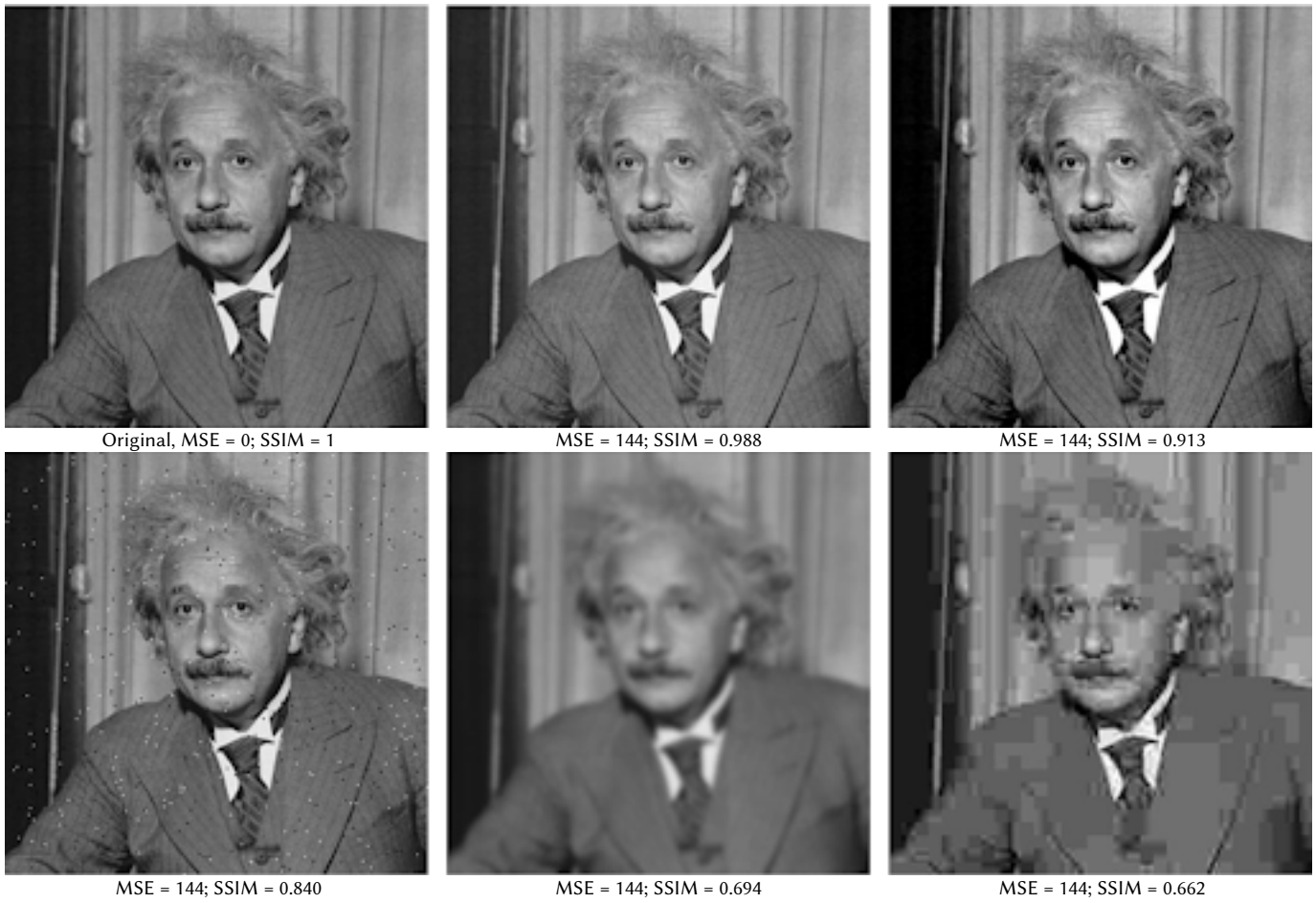


Fig. 2. Mean Squared Error (MSE) vs Structural Similarity Index Measure (SSIM).

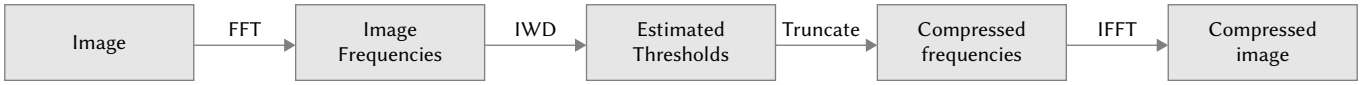


Fig. 3. Block diagram for image compression using Fast Fourier Transform (FFT) and Intelligent Water Drop Algorithm (IWD).

To obtain the compressed image in pixels, the Inverse Fourier Transform is carried out.

IV. Structural Similarity (SSIM) Index

The Structural Similarity (SSIM) index is a method employed to estimate the similarity in two images [18]. The SSIM index is studied as a measure of the quality of one of the images being examined, provided that the other image is considered to be of perfect quality. It is an enhanced version of the previously proposed universal image quality index.

For the examples shown in Fig. 2, all distorted images have about the same mean squared error (MSE) values for the initial image, but very distinct performance. SSIM offers a much better indication of image quality.

V. Our Proposed Solution

Using Fast Fourier Transforms, the compression of images involves thresholding the complex Fourier coefficients and applying reverse Fourier transform to the result in order to restore the image. These thresholds should be carefully chosen because the image can not be compressed by too low threshold while too high can result in very lossful compression. These thresholds are previously selected experimentally and are hard-coded for all images. Our experiments show that each compressed image tends to have its own set of thresholds, resulting in a better compression as well as a better quality ratio. But running the compression algorithm significantly for each image with different parameters is a very heavy and computationally expensive task to find an optimal solution. Also, the thresholds for RGB images are a triplet of 3 values instead of 1 value. Consequently, the complexity of such a task increases even more in the case of RGB images. There are some approximations available that can be used to estimate a set of threshold values, but none of them work well in terms of space or time complexity to our extent of knowledge.

For which, we suggest the intelligent water drops algorithm (IWD) to estimate these triplet's values. The task in hand is to estimate 3 discrete threshold values for each image compression. But producing these values directly may not be an optimal way to get the result. Instead, we try to obtain parameter (p1, p2, p3) for these 3 discrete values that are multiplied by the maximum absolute value of the RGB channel yield threshold of the complex fourier coefficients.

IWD optimizes several problem areas, such as n-queens, traveling salesman, multiple knapsack, but here we used IWD's traveling salesman variant to fulfill our task of finding the optimal three parameters as shown in Fig. 3.

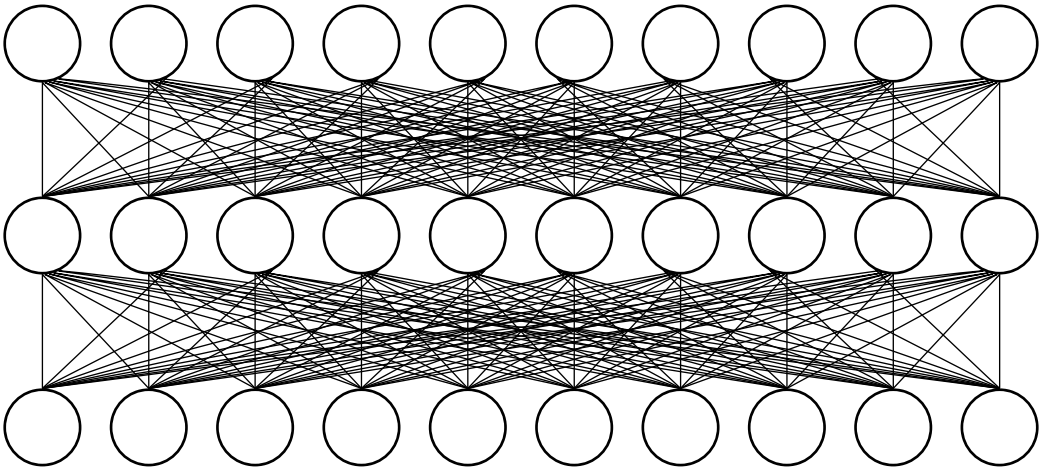


Fig. 4. Each layer has 10 nodes of values from 0.001 to 0.01 with step size of 0.001, where each node is connected to each other in the next layer. Layers 1, 2 and 3 may have p1, p2 and p3 values respectively. The IWD starts at layer 1 with any node and stops at layer 3 when it reaches a node.

(a) (b) (c) (d)

Fig. 5. (a) Sample image 1, (b) Sample image 2, (c) Sample image 3, (d) Sample image 4.

A. Modeling of the Problem

Our experiments with different threshold sets show that the value to be optimized for these parameters tends to lie in a limited region of parameter space, i.e. 0.001-0.01. We have modelled a graph of 30 nodes, 10 nodes for each of the 3 threshold triplet values by incorporating this prior knowledge about the parameter space. The value of each of the 10 nodes is 0.001 higher than the previous node (see Fig. 4). From these 10 discrete values, IWD is then used to find the best triplet combination that increases the image's SSIM along with the highest compression ratio.

The IWD algorithm computes the Local Heuristic H(j) for every node it moves forward to (j is the next node of the graph where the drop will move), for our task, we have used a constant heuristic 1.

Algorithm

- function quality:

1. input : values of $p1$, $p2$, $p3$
2. Calculate the maximum absolute value (say max_{val}) from fourier coefficients.
3. Calculate thresholds by multiplying the maximum absolute value with $p1$, $p2$ and $p3$.
4. $Thresholds = (p1 * max_{val}, p2 * max_{val}, p3 * max_{val})$
5. Filter the fourier coefficients by zeroing out those which are less than the calculated threshold.
6. Apply inverse fourier transform to the filtered coefficients to obtain decompressed image.
7. Calculate SSIM score using decompressed and original image.
8. return SSIM

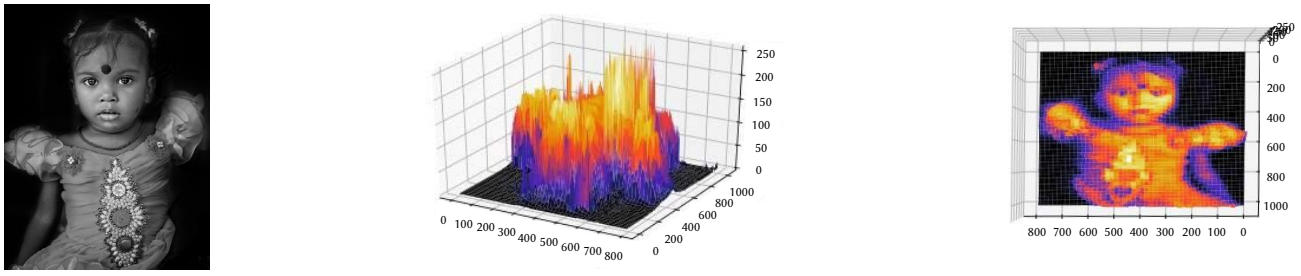
It is necessary to rank each solution produced by an IWD based on its quality. For each iteration, for each droplet and for the best global solution, IWD calculates this quality. We used a quality function for our task that takes the traveled path (in our case the p1, p2 and p3 values) and compresses the image using these thresholds and returns the SSIM score of that particular compression.

VI. Experimental Results

In simulations, to compare the results obtained from the traditional method and the IWD-based FFT compression method four sample images were selected. These images were directly scraped from the internet with query, "high fidelity images", "high resolution colored images" and "HD portraits". Fig. 5 shows the four sample images selected for experimental results.

Fig. 6, 8, 10 and 12 represent the grayscale, surface plot and top view plot of sample image 1, 2, 3 and 4 respectively. In the grayscale image, where the threshold is a single value, as applied in the image in Fig. 7, the visual quality of the image is preserved even after suppressing 98.17 per cent of the data in an image. Similarly in Fig. 9, Fig. 11, Fig. 13 the visual quality of the image is preserved even after the image is highly compressed. In the case of RGB images, the threshold is separate for each channel, thus the same procedure of hit and trial can be implemented for each channel separately.

It is concluded from Fig. 7, Fig. 9, Fig. 11, Fig. 13 that as more amount of data is suppressed, the less visually similar the images look. The quality of an image and the amount of compression is a tradeoff between each other. Hence, the threshold for compression must be chosen carefully, such that, the quality of the image along with compression maximizes. Every image has its own set of features along the channels that must be taken in order to achieve maximum quality.



(a) (b) (c)

Fig. 6. (a) represents the grayscale of sample image 1, (b) represents surface plot, and (c) represents top view plot.

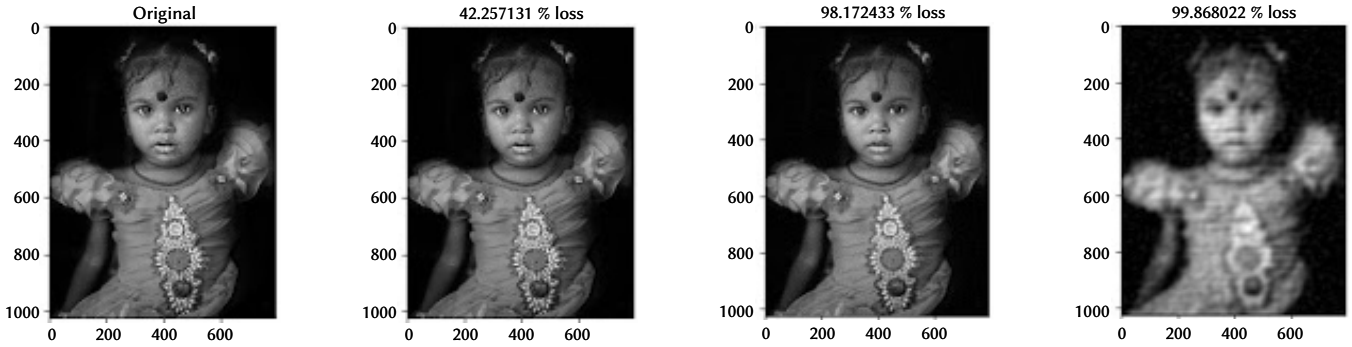
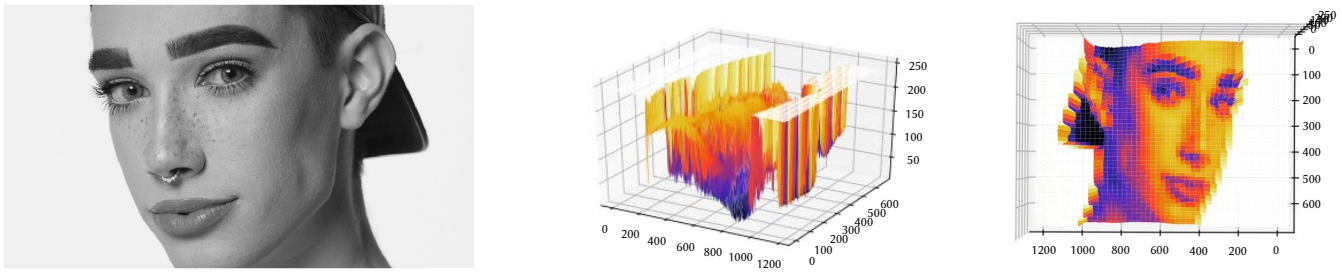


Fig. 7. Representations of the results of compression using Fourier transforms with the selected thresholds.



(a) (b) (c)

Fig. 8. (a) represents the grayscale of sample image 2, (b) represents surface plot, and (c) represents top view plot.

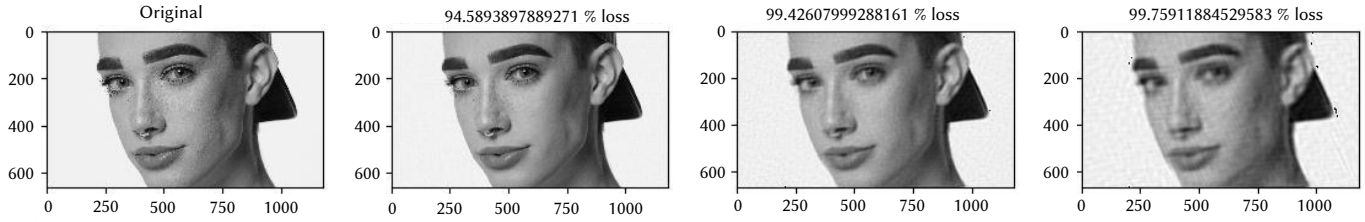
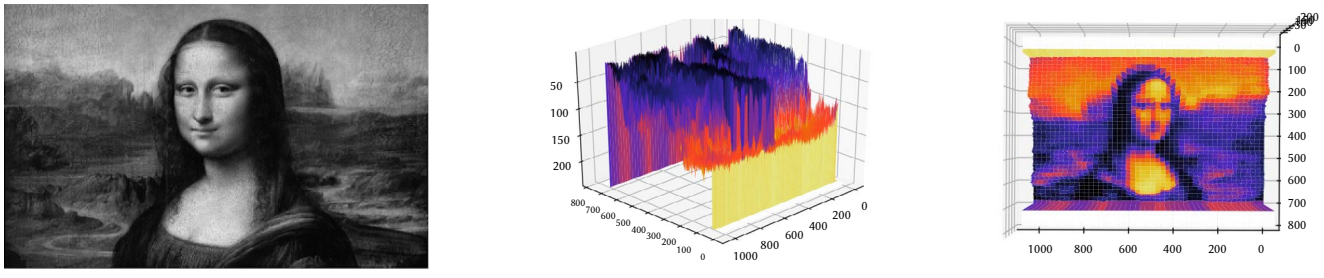


Fig. 9. Representations of the results of compression using Fourier transforms with the selected thresholds.



(a) (b) (c)

Fig. 10. (a) represents the grayscale of sample image 3, (b) represents surface plot, and (c) represents top view plot.

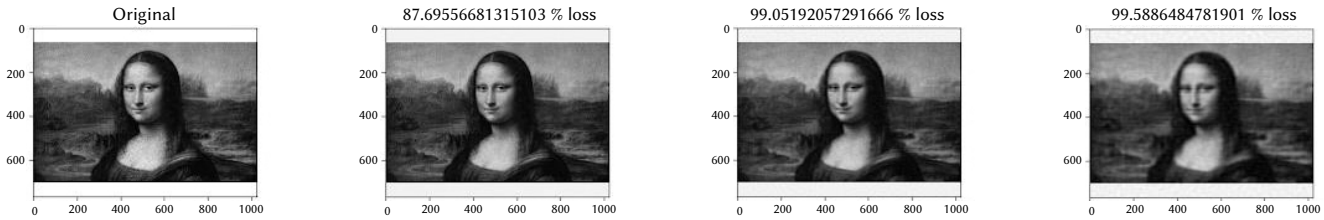
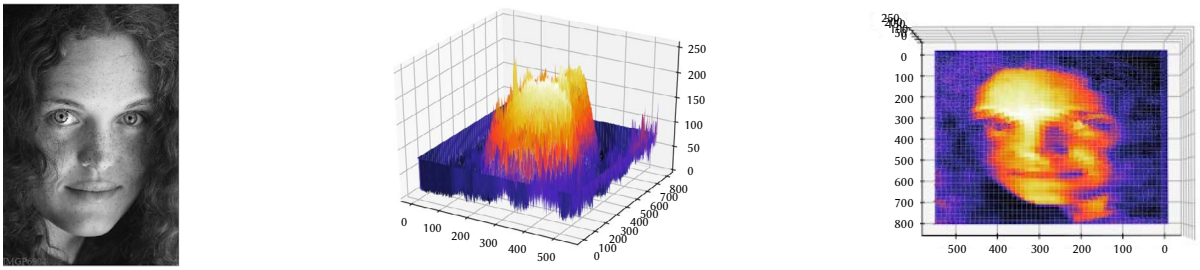


Fig. 11. Representations of the results of compression using Fourier transforms with the selected thresholds.



(a) (b) (c)

Fig. 12. (a) represents the grayscale image of sample image 4, (b) represents surface plot, and (c) represents top view plot.

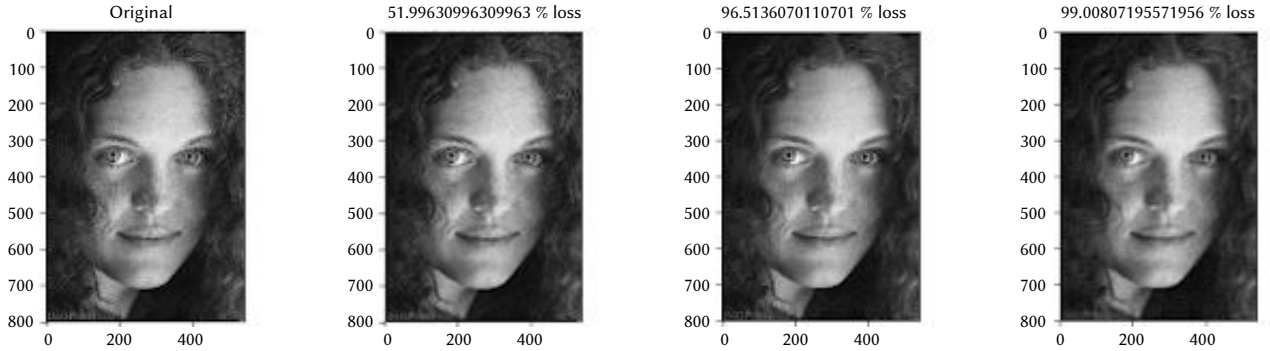


Fig. 13. Representations of the results of compression using Fourier transforms with the selected thresholds.

TABLE I. Data Loss and SSIM Score Comparison

Image	Compressed RGB image using threshold obtained using IWD		Compressed RGB image using standard method		Compressed RGB image using swapped values of threshold	
	SSIM Score	Data loss(%)	SSIM Score	Data loss(%)	SSIM Score	Data loss(%)
Sample image 1	0.69	99.8766642304	0.72	95.63583972020	0.60	95.63583972020
Sample image 2	0.73	99.7957887012	0.65	96.66549437568	0.65	95.95311462500
Sample image 3	0.88	99.7381844611	0.78	99.57618713378	0.69	96.58470153808
Sample image 4	0.62	98.9637915129	0.62	98.9637915129	0.65	79.35055350553

Hence, it is necessary to calculate thresholds based on the content of the image.

Table I shows the results of compressing images using the defined values and calculated values of thresholds for 4 different images of different color densities and structural formations. For sample image 1, using predefined values of (0.001, 0.001, 0.001) as p1, p2 and p3, the SSIM score after compression was found out to be 0.72 while the amount of data loss was over 95%. While using the values (0.001, 0.001, 0.01) which were obtained by IWD, the SSIM score was found out to be 0.69 while achieving a compression loss of more than 99.8%. To get a clear understanding of the result, we switched the values of the triplets to (0.01, 0.01, 0.001) and performed a compression. The image obtained had a SSIM score of 0.60 along with a 95% data loss. These results explain that, to achieve a maximum compression along with good quality for this particular image, the green channel must be kept along with blue being the least required. Which when violated, the result became unstable (see Table I: sample image 1).

VII. Conclusion

Image Compression which is the science of reducing the amount of data required to represent an image, is one of the most useful and commercially successful technologies in the field of digital image processing. Our assessment demonstrates that each compressed image must integrate the image content, thus improving the need to evaluate compression parameters for each image and whereby thresholds must be calculated for each image using IWD to compress images using fourier transform. Here we have constrained the values

of these thresholds to be in a bounded parameter space. In future implementations, we wish to overcome the restricted parameter space implementation, i.e, the values that we used to begin the IWD search. We believe that, if an elaborated search space aka parameter space for the threshold values is provided to the IWD during its initialization, a better minimum can be found for our loss function thus improving the optimization strategy. Hence, we would like to introduce a bigger parameter space for finding the threshold values algorithm so as to increase the efficiency of the methodology and achieve better results.

References

[1] J. R. Jensen, *Introductory digital image processing: a remote sensing perspective,* United States: N. p., 1986, Web.

[2] Y. Wang, J. Ostermann, and Y. Q. Zhang, *Video processing and communications* (Vol. 1). Upper Saddle River, NJ: Prentice hall. 2002.

[3] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte. "Breast cancer histopathological image classification using convolutional neural networks," In *IEEE 2016 international joint conference on neural networks(IJCNN),* Vancouver, BC, Canada 2016, July, pp. 2560-2567.

[4] R. Hartley, and A. Zisserman. *Multiple view geometry in computer vision.* Cambridge university press, 2003.

[5] L. R. Long, L. E. Berman, L. Neve, G. Roy, and G.R. Thoma, March. "Application-level technique for faster transmission of large images on the internet," *Proc. SPIE 2417, Multimedia Computing and Networking* 1995, (14 March 1995); https://doi.org/10.1117/12.206077

[6] D. Taubman, and M. Marcellin, *JPEG2000 image compression fundamentals, standards and practice: image compression fundamentals, standards and practice* Springer Science Business Media, vol. 642, 2012.

[7] B. Gupta, M. Tiwari, and S. S. Lamba. "Visibility improvement and mass segmentation of mammogram images using quantile separated histogram equalisation with local contrast enhancement," *CAAI Transactions on Intelligence Technology* vol. 4, no. 2, pp. 73-79, 2019.

[8] S. Ghosh, et al. "Graphology based handwritten character analysis for human behaviour identification," *CAAI Transactions on Intelligence Technology* vol. 5, no. 1, pp. 55-65, 2020.

[9] W. Eng, V. Koo, and T. Lim. "IPDDF: an improved precision dense descriptor based flow estimation," *CAAI Transactions on Intelligence Technology*, vol. 5, no. 1, pp. 49-54, 2020.

[10] D. Taubman, "High performance scalable image compression with EBCOT," *IEEE Transactions on image processing,* vol. 9, no. 7, pp. 1158-1170, 2000.

[11] A. S. Lewis, and G. Knowles, "Image compression using the 2-D wavelet transform," *IEEE Transactions on image Processing,* vol. 1, no. 2, pp. 244-250, 1992.

[12] A. Said, and W. A. Pearlman, "An image multiresolution representation for lossless and lossy compression," *IEEE Transactions on image processing,* vol. 5, no. 9, pp. 1303-1310, 1996.

[13] Z. Guo, Y. Shen, A. K. Bashir, M. Imran, N. Kumar, D. Zhang and K. Yu, "Robust Spammer Detection Using Collaborative Neural Network in Internet of Thing Applications," IEEE Internet of Things Journal, vol. 8, no. 12, pp. 9549-9558, 15 June 2021, doi: 10.1109/JIOT.2020.3003802.

[14] A. Hore, and D. Ziou, "Image quality metrics: PSNR vs. SSIM," In 2010 20th International Conference on Pattern Recognition 2010, August. (pp. 2366-2369). IEEE.

[15] J. Hu, J. Deng, and J. Wu, "Image compression based on improved FFT algorithm," *Journal of Networks*, vol. 6, no. 7, pp.1041-1048, 2011.

[16] H. Shah-Hosseini, "The intelligent water drops algorithm: a nature-inspired swarm-based optimization algorithm," *International Journal of Bio-inspired computation,* vol. 1, no. 1-2, pp. 71-79, 2009.

[17] K. Yu, L. Tan, L. Lin, X. Cheng, Z. Yi and T. Sato, "Deep-Learning-Empowered Breast Cancer Auxiliary Diagnosis for 5GB Remote E-Health," *IEEE Wireless Communications*, vol. 28, no. 3, pp. 54-61, June 2021, doi: 10.1109/MWC.001.2000374.

[18] Z. Wang, A. C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE transactions on image processing, vol. 13, no. 4, pp. 600-612, 2004.

[19] S. D. Kamble, N. V. Thakur, and P. R. Bajaj. "Modified Three-Step Search Block Matching Motion Estimation and Weighted Finite Automata based Fractal Video Compression," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 4, 2017.

[20] L. E. George, and H. A. Hadi. "User Identification and Verification from a Pair of Simultaneous EEG Channels Using Transform Based Features," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 5, pp. 54-62, 2019.

[21] F. López, L. de la Fuente Valentín, and Í. S. M. de Mendivil. "Detecting image brush editing using the discarded coefficients and intentions," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 5, pp. 15-21, 2019.

[22] N. V. Malathkar, S. K. Soni, "High compression efficiency image compression algorithm based on subsampling for capsule endoscopy," Multimedia Tools and Applications, vol. 80, pp. 22163–22175, 2021, doi: 10.1007/s11042-021-10808-0.

[23] R. D. Sivakumar, K. R. Soundar, "A novel generative adversarial block truncation coding schemes for high rated image compression on E-learning resource environment," Materials Today: Proceedings (2021), ISSN 2214-7853, doi: 10.1016/j.matpr.2021.01.270.

Surinder Kaur

Ms. Surinder Kau is currently working as an Assistant Professor in Bharati Vidyapeeth's College of Engineering, Guru Gobind Singh Indraprastha University, Delhi, India. She received the B.E. degree in Computer Science and Engineering in 2000 from Sant Longowal Institute of Engineering and Technology, Longowal, Punjab, and the M.Tech. degree in Informaton Technology from Punjabi University, Patiala, Punjab, India, in 2003. Her current research interests include soft computing, Biometrics, Artificial Intelligence.

Gopal Chaudhary

Dr. Gopal Chaudhary is currently working as an assistant professor in Bharati Vidyapeeth's College of Engineering, Guru Gobind Singh Indraprastha University, Delhi, India. He holds a Ph.D. in Biometrics at the division of Instrumentation and Control engineering, Netaji Subhas Institute of Technology, University of Delhi, India. He received the B.E. degree in electronics and communication engineering in 2009 and the M.Tech. degree in Microwave and optical communication from Delhi Technological University (formerly known as Delhi College of Engineering), New Delhi, India, in 2012. He has 50 publications in refereed National/International Journals & Conferences (Elsevier, Springer, Inderscience) in the area of Biometrics and its applications. His current research interests include soft computing, intelligent systems, 8 information fusion and pattern recognition. He has organized many conferences and leading special issues in Taylor and Francis, Springer, IOS Press, Computers Materials & Continua etc.

Dinesh Kumar

Dr. Dinesh Kumar is working as assistant professor in the Department of Electronics communication Engineering, Lingaya's University, Haryana. He holds a Ph.D. in Electronics and Communication Engineering Specialized in VLSI through IOT from Lingaya's Vidyapeeth (Deemed to be University), Faridabad, Haryana in 2019. He received the B.E. degree in electronics and communication engineering in 2012 and the M.Tech. degree in VLSI System Design from Jawaharlal Nehru Technological University, Hyderabad in 2014. He has several publications like journals and conferences in different national and international publishing houses. He has guided several M.Tech and B.Tech students in their dissertation and major minor projects as well as supervising Ph.D. Scholars. His area of research interest is Internet of things (IOT), Cloud Computing, VLSI through IOT, VLSI through Optical communication, VLSI through Embedded systems etc.

Manu S Pillai

Mr. Manu S Pillai is currently working as an Instructor and Product Engineer for Machine Learning and Data Science in Coding Blocks, New Delhi, India. He is also associated with Information Technology Department, Bharati Vidyapeeth's College of Engineering, Guru Gobind Singh Indraprastha University, Delhi, India. His current research interests include Computer Vision, Image Processing & Deep Learning.

Yash Gupta

Mr. Yash Gupta is student of Bharti Vidyapeeth's College of Engineering (GGSIPU), New Delhi is currently pursuing his bachelor's degree in Information Technology (2016-2020). His all-round abilities of working in both the domains and knack of learning something new always, motivated him to work in many projects and gaining technical as well as non-technical expertise. He has received many awards and certifications from multiple organizations regarding the same. He has shown keen interest in Cloud Computing, Machine Learning, Financial Analysis and Full-Stack Development.

Manju Khari

Dr. Manju Khari is an Associate Professor in School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, India. She was 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, information security, optimization, Image processing and machine learning. She has 70 published papers in refereed National/International Journals & Conferences (viz. IEEE, ACM, Springer, Inderscience, and Elsevier) and 10+ edited books from reputed publishers. She is also co-author of two books published by NCERT of Secondary and senior Secondary School.

Vicente García-Díaz

Prof. Vicente García-Díaz is an Associate Professor in the Department of Computer Science at the University of Oviedo (Languages and Computer Systems area). He earned a PhD in Computer Science from the University of Oviedo as well as degrees in Computer Engineering and Technical Systems Computer Engineering. In addition, he earned a degree in Occupational Risk Prevention. He has supervised 100+ academic projects, he has published 100+ research papers in journals, conferences and books from prestigious publishers. He is also a member of the editorial and advisory boards of several journals and books. His research interests include eHealth, eLearning, machine learning and the use of domain specific languages in different areas.

Javier Parra Fuente

Prof. Javier Parra Fuente is a full-time professor at the Universidad Internacional de La Rioja (UNIR). His current research interests are in software and web engineering, artificial intelligence and public administration services. He is deputy director of the international academic development department at UNIR. In addition, he is information and communications technology officer of the government of Spain. He has been professor in several universities during more than 20 years: Universidad Autónoma de Madrid, Marconi International University, Universidad Pontificia de Salamanca. He was a postdoctoral researcher at the Oxford University during 2 years and visiting professor at seven universities in Argentina, Colombia, the Dominican Republic, Mexico and Peru. Director of several university master's degrees and professor in different university doctoral, master and degree programs.

Modeling Sub-Band Information Through Discrete Wavelet Transform to Improve Intelligibility Assessment of Dysarthric Speech

Laxmi Priya Sahu[1], Gayadhar Pradhan[1], Jyoti Prakash Singh[2] *

[1] Department of Electronics and Communication Engineering, National Institute of Technology Patna, (India)
[2] Department of Computer Science and Engineering, National Institute of Technology Patna (India)

Received 26 January 2022 | Accepted 25 March 2022 | Published 3 October 2022

Abstract

The speech signal within a sub-band varies at a fine level depending on the type, and level of dysarthria. The Mel-frequency filterbank used in the computation process of cepstral coefficients smoothed out this fine level information in the higher frequency regions due to the larger bandwidth of filters. To capture the sub-band information, in this paper, four-level discrete wavelet transform (DWT) decomposition is firstly performed to decompose the input speech signal into approximation and detail coefficients, respectively, at each level. For a particular input speech signal, five speech signals representing different sub-bands are then reconstructed using inverse DWT (IDWT). The log filterbank energies are computed by analyzing the short-term discrete Fourier transform magnitude spectra of each reconstructed speech using a 30-channel Mel-filterbank. For each analysis frame, the log filterbank energies obtained across all reconstructed speech signals are pooled together, and discrete cosine transform is performed to represent the cepstral feature, here termed as discrete wavelet transform reconstructed (DWTR)- Mel frequency cepstral coefficient (MFCC). The *i-vector* based dysarthric level assessment system developed on the universal access speech corpus shows that the proposed DTWR-MFCC feature outperforms the conventional MFCC and several other cepstral features reported for a similar task. The usages of DWTR-MFCC improve the detection accuracy rate (DAR) of the dysarthric level assessment system in the text and the speaker-independent test case to 60.094 % from 56.646 % MFCC baseline. Further analysis of the confusion matrices shows that confusion among different dysarthric classes is quite different for MFCC and DWTR-MFCC features. Motivated by this observation, a two-stage classification approach employing discriminating power of both kinds of features is proposed to improve the overall performance of the developed dysarthric level assessment system. The two-stage classification scheme further improves the DAR to 65.813 % in the text and speaker-independent test case.

Keywords

Approximation Coefficient, Cepstral Coefficients, Detail Coefficient, DWT, Dysarthria, Dysarthric Level, IDWT, Sub-Band Signal.

DOI: 10.9781/ijimai.2022.10.003

I. Introduction

Dysarthria reduces the speech intelligibility of a person by affecting the speech production system [1]. Parkinson's disease, amyotrophic lateral sclerosis, cerebral palsy, brain tumor, and brain injury are some of the causes for developing dysarthria in a person [2]–[4]. The speech intelligibility of a dysarthric person varies from near-normal to unintelligible depending on the level of severity. From the mid-level of severity, it is difficult to understand the spoken utterances of a dysarthric person by unfamiliar listeners [5]. In the conventional approach, the intelligibility level of a spoken utterance is measured by subjective assessment in clinical applications. The subjective assessment approach is costlier in time and money with a possibility of biasness towards the previous knowledge of experts about the type of disease [6]. Due to easy accessibility and consistent performance, recently, automated objective assessment methods are explored for the diagnosis of dysarthria in the primary stages [7].

The speech quality of a person suffering from dysarthria differs from a normal speaker due to change in loudness level, fundamental frequency, voice instability, voice breaks, and speaking rate [8], [9]. Consequently, the performance of the automatic speech recognition (ASR) system for a dysarthric speaker degrades compared to the normal speaker [10]. Using this aspect, some of the reported works used the ASR system for evaluating the level of dysarthria [10]–[14]. The word recognition rate (WRR), state-level log-likelihood ratios (SLLRs), and log-likelihoods (LLs) are used as the measuring parameter for the evaluation of the intelligibility level of the dysarthric speech. Such approaches are more suitable when a fixed set of words are used for testing different speakers. The performance of the ASR system also varies depending on the linguistic context of the speaker. The scarcity

* Corresponding author.
E-mail addresses: laxmipriya.ec16@nitp.ac.in (L. P. Sahu), gdp@nitp.ac.in (G. Pradhan), jps@nitp.ac.in (J. P. Singh).

of dysarthric speech data and limited availability of reference language models make the blind intelligibility assessment methods more useful [15]. The blind intelligibility assessment approaches use a classifier for differentiating the healthy speaker and dysarthric speaker and further separate according to their severity level. Some of the blind modeling approaches include the classical modeling methods like support vector machine (SVM) [16], Gaussian mixture model(GMM) [17], and recently reported neural network (NN) based modeling methods [18]–[22]. The NN based methods such as artificial neural network (ANN) [18], deep neural network (DNN) [19], convolutional neural network (CNN) [23], [24], long short-term memory network (LSTM) [25], bidirectional LSTM (BLSTM) and recurrent neural network (RNN) [26] have been explored for the intelligibility assessment of dysarthric speech. The *i-vector* representation of input speech data is also explored for assessment of dysarthria [27], [28]. The *i-vector* based representation maps the varying length of input speech utterance into a fixed dimension. Various feature projection and scoring schemes in combination with *i-vector* improve the performance of dysarthric level assessment system [27], [28]. Some of the reported works also used the combination of various statistical modeling and NN-based approaches [16], [17], [29]–[31]. Despite the use of sophisticated acoustic modeling methods the performances reported for dysarthric level assessment are less.

The aforementioned modeling approaches mostly use the spectral domain features for acoustic representation of the input speech data [18], [32]. The spectral representation of the input speech data such as spectrogram [33], log filterbank energy [26], Mel-frequency cepstral coefficients (MFCCs) [32], multitaper MFCC [18], perceptual linear prediction cepstral coefficients (PLPCCs) [34], [35], linear prediction cepstral coefficients (LPCC) [36], constant Q cepstral coefficients (CQCCs) [37] and line spectral frequencies have been explored for the assessment of dysarthric level. Several voice quality features [38]–[40], prosodic features [17], and excitation source features [41]–[43] are also explored for the assessment of dysarthria. The extraction of voice quality and prosodic feature from a speech signal is difficult and performance is highly dependent on the employed feature extraction approach. On the other hand, the cepstral feature can be easily extracted from the input speech data and very frequently used in the development of speech based applications.

The speech signal within a sub-band varies at a fine level depending on the type and level of dysarthria. Following the human perception of the speech signal [32], most of the cepstral features are extracted by analyzing the short-term magnitude spectra using a Mel-filterbank. Consequently, the fine level information present in the higher frequency regions is smoothed out due to the larger bandwidth of Mel-filters. The discriminating information present at the fine level in the short-term magnitude spectra can be captured up to a certain level by increasing the size of the filterbank. The increase in filterbank size may also capture the redundancy present in the lower frequency regions. Alternatively, the fine level information can be captured by decomposing the speech signal into different sub-band signals and analyzing the magnitude spectra of each sub-band signal [14]. Motivated by these observations, in this paper, we have firstly decomposed the speech signal using discrete wavelet transform (DWT) [44] into approximation and detail coefficients, respectively, at each level. The speech signals representing different sub-bands are then reconstructed using inverse DWT (IDWT) [44], [45]. In the process of IDWT, the speech signals are reconstructed by using the detail coefficient obtained at each level of decomposition and making all other coefficients to zero vector. Finally, at the last level of decomposition, the speech signal representing the lower frequency region is reconstructed by using only the approximation coefficients and making all detail coefficients to zero vectors. The log filterbank energies are computed by analyzing the short-term discrete Fourier transform (DFT) magnitude spectra of each reconstructed speech using the Mel-filterbank. For each analysis frame, the log filterbank energies obtained across all reconstructed speech are pooled together, and discrete cosine transform (DCT) is applied to represent the 13-dimensional base cepstral feature, here termed as discrete wavelet transform reconstructed - Mel frequency cepstral coefficient (DWTR-MFCC). The experimental results presented in this study show that DWTR-MFCC enhances discrimination among the overlapping classes compared to the conventional MFCC feature [46]. It also carries additional information to MFCC features. Motivated by this observation finally, a two-stage classification scheme is proposed to improve the overall performance of the dysarthric assessment system.

The remainder of the paper is organized as follows: Section II describes the proposed feature extraction method using DWT for the assessment of dysarthric level. Section III presents experimental setup for the development of *i-vector* based dysarthric level assessment system. The experimental results are presented in Section IV. Finally, Section V concludes this study.

II. Proposed Feature for Assessment of Dysarthric Level

The wavelet transform-based approaches are most preferred for time-frequency analysis of different types of signals [44], [47]–[50]. In the following section, decomposition and reconstruction of input speech using DWT followed by the proposed approach for the computation of DWTR-MFCC feature are presented.

A. Decomposition and Reconstruction of Speech Signal Using DWT

Using DWT, the speech signal $s(n)$ can be decomposed into high-frequency detail coefficients and low-frequency approximation coefficients by passing through a series of high-pass and low-pass filters, respectively. At a particular level of decomposition, the detail coefficient ($D_{i,j}$), and approximation coefficient ($A_{i,j}$) can be obtained as given in [44]. Equation (1) and Equation (2) represents the detail coefficients and approximation coefficients of the input signal $s(n)$, respectively.

$$D_{i,j} = \sum_{n} s(n)\psi_{i,j}(n) \tag{1}$$

$$A_{i,j} = \sum_{n} s(n)\phi_{i,j}(n) \tag{2}$$

where integer i and j provide the information about the amount of scaling and shifting of the wavelet function, respectively. The mother wavelet function ($\psi_{i,j}(n)$) and the father wavelet function ($\phi_{i,j}(n)$) are extracted from the continuous wavelet transform (CWT) using most commonly used dyadic grid arrangement [44], [47]. Equation (3) represents the mother wavelet function ($\psi_{i,j}(n)$).

$$\psi_{i,j}(n) = 2^{-i/2}\psi\left(2^{-i}n - j\right) \tag{3}$$

Equation (4) represents the father wavelet function ($\phi_{i,j}(n)$).

$$\phi_{i,j}(n) = 2^{-i/2}\phi\left(2^{-i}n - j\right) \tag{4}$$

Equation (5) refers to the representation of the signal $s(n)$ as the combination of detail and approximation coefficients.

$$s(n) = \sum_{j=-\infty}^{\infty} A_{i_0,j}\phi_{i_0,j}(n) + \sum_{i=-\infty}^{i_0}\sum_{j=-\infty}^{\infty} D_{i,j}\psi_{i,j}(n) \tag{5}$$

where, $A_{i_0,j}$ represents the approximation coefficient at level i_0. The approximation coefficient at i^{th} level of decomposition can be obtained by combining the detail and approximation coefficients obtained at

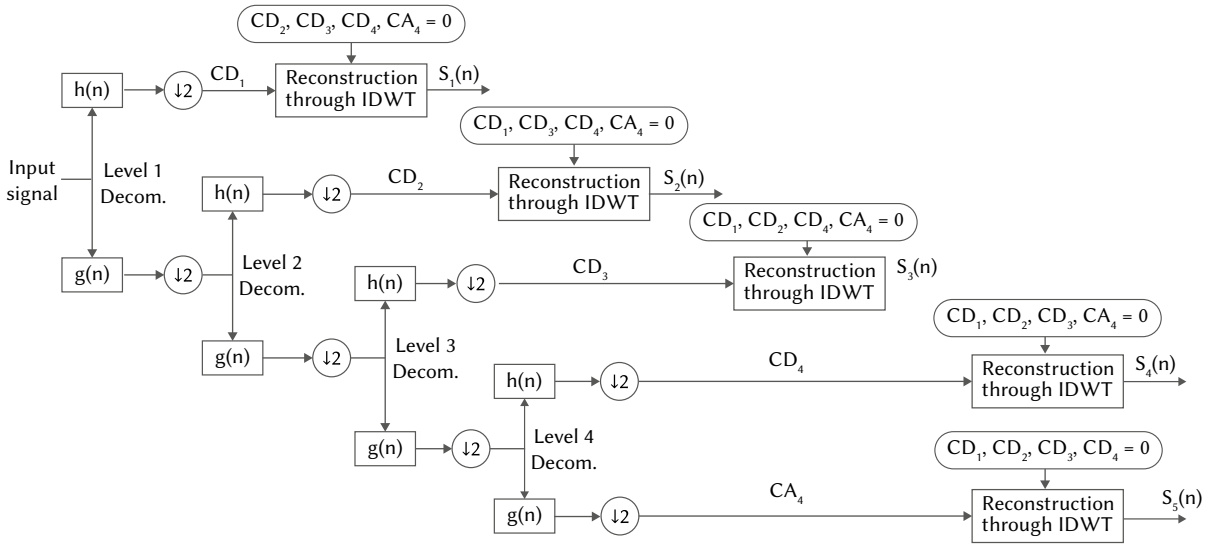


Fig. 1. The block diagram illustrating the DWT based four level multiresolution decomposition of the speech signal. The $h(n)$ and $g(n)$ represents the high pass and lowpass filter, respectively. CD_1, CD_2, CD_3, CD_4 and CA_4 represents the detail coefficient at level1, level2, level3, level4 and approximation coefficient at level4 decomposition, respectively. $s_l(n)$, $l = 1; 2; ...5$ represents the reconstructed signals.

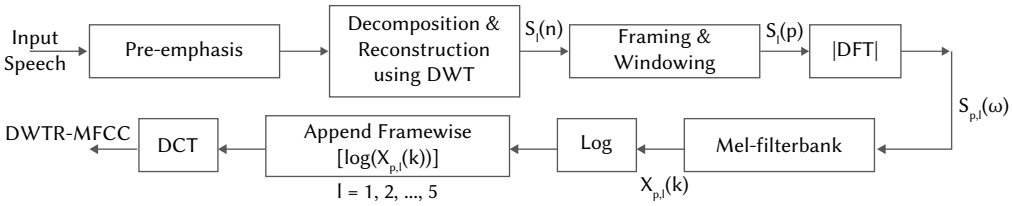


Fig. 2. The block diagram representation of proposed DWTR-MFCC feature extraction process.

$(i+1)^{th}$ decomposition level [44]. The block diagram representing four-level of wavelet decomposition and reconstruction of five sub-band speech signals is illustrated in Fig. 1. Let, the original speech signal $s(n)$ be the approximation coefficient at level zero. As shown in the block diagram, in each level, the approximation coefficient is decomposed into detail and approximation coefficients of the next higher level. The decomposition is performed by processing the approximation coefficient through a pair of high-pass and low-pass filters having impulse response $h(n)$ and $g(n)$, respectively. At each level, decomposed signals are downsampled to half of the original sampled signal ($\downarrow 2$) to remove the redundant samples while satisfying Nyquist's criteria [44], [47]. As shown in the block diagram, a four-level DWT-based decomposition finally results in four detail coefficients (CD_1, CD_2, CD_3 and CD_4) and one approximation coefficient (CA_4). The five sub-band signals are then reconstructed through inverse DWT (IDWT) [44],[50]. During the reconstruction process, one coefficient is preserved while the other coefficients are made to be a zero vector. On the use of one coefficient vector in the IDWT reconstruction process, the resulting signal contains frequency components only in that region. Therefore, the reconstructed sub-band speech $s_l(n), l = 1, 2, ... 5$ represents 4−8 kHz, 2−4 kHz, 1−2 kHz, 0.5−1 kHz, and 0−0.5 Hz frequency band, respectively for a 16kHz sampled speech data.

B. Computation of Proposed DWTR-MFCC Feature

The proposed method for the computation of the DWTR- MFCC feature is depicted in Fig. 2. As shown in the block diagram, the input speech signal $s(n)$ is processed through the following sequence of steps to extract the DWTR-MFCC feature.

1. The speech signal is subject to a pre-emphasis filter with a filter coefficient of 0.97 to boost the high-frequency component. As explained in the previous section, the decomposition and reconstruction of five sub-band speech signals are then performed using the Daubechies (db) wavelet function [45]. Let, the reconstructed sub-band signals are represented by $s_l(n), l = 1, 2, ... 5$.
2. The short-term analysis of each reconstructed signal $s_l(p)$ is performed by processing with a fixed-length Hamming window of size 20 ms with a frame-shift of 5 ms. The short-term magnitude spectra are then computed by performing DFT on the short-term analysis frames $s_l(p)$, where $p = 1, 2, ... P$. The total number of analysis frames is represented by P. The short-term DFT magnitude spectra for p^{th} analysis frame of l^{th} sub-band signal is denoted by $S_{p,l}(\omega)$. The nature of short-term magnitude spectra for original and reconstructed sub-band signals are depicted in Fig. 3. This analysis is performed for a center frame of vowel /a/ taken from the dysarthric speaker. The logarithmically compressed magnitude spectrum of the original frame is given in Fig. 3 (a). The logarithmically compressed magnitude spectra obtained for sub-band signals reconstructed using CA_4, CD_4, CD_3, CD_2 and CD_1 are given in Fig. 3(b)-Fig. 3(f), respectively. The magnitude spectra for each sub-band signal are different. Therefore, fine level information can be extracted by analyzing the spectra separately.
3. The Mel-frequency warping of each short-term magnitude spectra $S_{p,l}(\omega)$ is performed using a 30 channel Mel-filters. The size of the Mel-filterbank has remained the same for each sub-band signal $s_l(n)$. The filterbank energies are then computed by following the standard procedure of the MFCC feature extraction. Here, the Mel-filterbank energy for the p^{th} analysis frame of l^{th} sub-band signal is represented by $X_{p,l}(k)$.
4. The Mel-filterbank energies are logarithmically compressed to reduce the dynamic range. For each analysis frame, log compressed filterbank energies obtained across all the sub-band signals are pooled together and discrete cosine transform (DCT) is performed to compute the 13-dimensional base DWTR-MFCC feature.

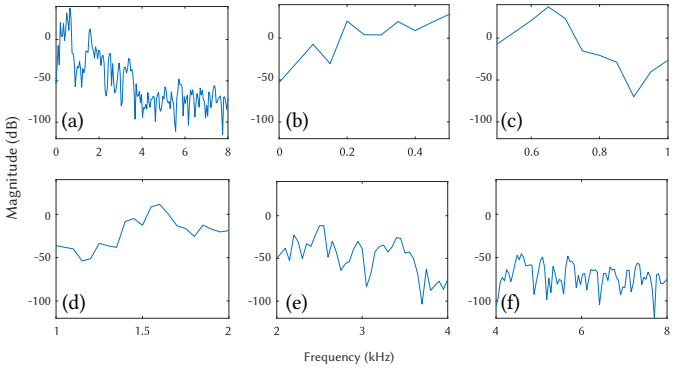


Fig. 3. Plot illustrating the nature of DFT magnitude spectra for sub-band signals. This analysis is performed for a center frame of vowel {a{ taken from the dysarthric speaker M11 of UA-speech corpus. (a) Logarithically compressed magnitude spectrum of the original frame. (b)-(f) Logarithmically compressed magnitude spectra obtained for sub-band signals reconstructed using CA_4, CD_4, CD_3, CD_2 and CD_1, respectively.

III. Experimental Setup

The automatic assessment of dysarthric level is performed using universal access speech (UA-Speech) corpus [51]. This database contains the speech data from 15 dysarthric speakers, which includes 4 female and 11 male speakers. The speech data of each speaker contains isolated words in three different blocks (B_1, B_2, and B_3). Each block contains a total of 255 words out of which 155 words are repeated and the rest 100 words are uncommon for all blocks. The phoneme level diversity is present in the database due to the availability of monosyllables, bisyllables, and polysyllables with the combination of various words. The subjective intelligibility of the speakers varies from 2% to 95%, which is classified into four intelligibility levels, namely very-low (0%-25%), low (26%-50%), mid (51%-75%) and high (76%-100%).

In this study, four datasets namely, the training dataset (*Train_Data*), development dataset (*Dev_Data*), speaker-dependent test dataset (*SD_Test*), and speaker-independent test dataset (*SI_Test*) are derived from the UA-Speech corpus. These datasets are prepared by balancing the recording microphone to remove the sensor effect. The selection of speaker and utterances for each dataset is done as follows:

1. *Train_Data*: This dataset contains the common words from all the blocks and uncommon words from two blocks (B_1 and B_3) of 8 speakers.
2. *Dev_Data*: This dataset contains a part of *Train_Data* equally distributed among dysarthric levels, speakers, and microphones, which is used for the development of universal background model (UBM), total variability matrix (T-matrix) and learning matrices for projection of *i-vectors* to lower dimensional subspace.
3. *SD_Test*: This dataset contains the speech utterances of the 100 uncommon words from block B_2 of 8 speakers present in the *Train_Data*. This dataset is speaker-dependent and text-independent.
4. *SI_Test*: This dataset contains speech data of 100 uncommon words from block B_2 of the remaining 7 speakers. Therefore, this dataset is speaker and text-independent compared to *Train_Data*.

The speakers selected for each dataset is summarized in Table I. All the experimental studies are performed using 16 kHz sampled speech data.

TABLE I. Speakers Selected for Training Dataset (*Train_Data*), Development Dataset (*Dev_Data*), Speaker-Dependent Test Dataset (*SD_Test*) and Speaker-Independent Test Dataset (*SI_Test*). The Abbreviations F and M Refer to the Female and Male Speakers, Respectively

Intelligibility Level	Train_Data, Dev_Data and SD_Test	SI_Test
Very low(0%-25%)	F03, M04	M12, M01
Low(26%-50%)	F02, M07	M16
Mid(51%-75%)	F04, M05	M11
High(76%-100%)	F05, M08	M09, M10, M14

A. Front-End Speech Parameterization

In all the experimental studies, the speech data is firstly pre-emphasized using a high-pass filter with a filter coefficient of 0.97. The short-term analysis of the pre-emphasized speech signal is performed by using an overlapping Hamming window of 20 ms duration at a frame-shift of 5 ms. The performance of dysarthric level assessment system employing proposed DWTR-MFCC feature is compared with MFCC [46], [52], linear prediction cepstral coefficient (LPCC) [53], constant Q cepstral coefficients (CQCCs) [37] and spectral moment time-frequency distribution augmented by low-order cepstra (SMAC) [54] acoustic features. The dimension of base DWTR-MFCC, MFCC, LPCC and CQCC is fixed at 13. As presented in [54], [55], the dimension for the base SMAC feature is fixed at 18, which contains 16 first-order spectral moments extracted using 16-channel Mel-spaced Gabor filterbank and two first-order coefficients of Garbor filtered spectra. To further analyze the impact of shape and size of filterbank on MFCC, the features are extracted using the varied size of Mel and linear filterbank. However, for each case, the 12 dimensional cepstral coefficients along with the energy coefficients are used as the feature vector [28], [46]. For each acoustic feature, the delta (Δ) and delta-delta ($\Delta - \Delta$) coefficients are computed using two preceding and two succeeding feature vectors from the current feature vector. The base feature is then appended to delta (Δ) and delta-delta ($\Delta - \Delta$) features. The feature vectors corresponding to the non-speech regions are removed by processing the speech signal through an energy-based voice activity detection (VAD) [56]. The cepstral mean-variance normalization (CMVN) [57] are then applied to the selected feature vectors to follow a zero mean unit variance distribution.

B. Development of I-Vector Based Dysarthric Assessment System

In the *i-vector* based approach, for the given set of acoustic feature vectors, a lower-dimensional vector of fixed size is created to represent the input speech data [58]–[60]. In this approach, firstly, the class-dependent Gaussian mixture model (GMM) mean supervector is created by adapting a class independent universal background model (UBM). The GMM mean supervector is then projected into a lower-dimensional subspace for mapping the given utterances to a fixed dimension, as proposed in [58]. Equation (6) refers to the *i-vector* representation of a given speech utterance.

$$M = m + Tw \tag{6}$$

where M is the adapted GMM mean supervector, m is the UBM mean supervector, T is the total variability matrix and w is the *i-vector*.

The *i-vectors* extracted from a given speech utterance contains speaker and sensor information along with the information of dysarthria. Therefore, for improving the performance of the dysarthric assessment system, the speaker and channel information need to be normalized. In this study, we have explored the linear discriminant analysis (LDA) [61] and within-class covariance normalization (WCCN) [62] for reducing the session and channel variabilities. The performance of the developed dysarthric assessment system is also evaluated by applying WCCN to the dimensionality reduced *i-vectors* obtained through LDA. The four levels dysarthric assessment is performed by comparing the *i-vectors* of the test speech with the trained *i-vectors* of each dysarthric level. We have performed both

cosine kernel [63] and probabilistic linear discriminant analysis (PLDA) [64] based scoring mechanisms. The class representative *i-vectors* for a particular dysarthric level is created by pooling all the speech data corresponding to that class. During testing, the *i-vectors* are extracted from each test data and compared with the trained *i-vector* of each class. The assignment of the test data to a particular class is done based on the maximum score. In this study, UBM model contains 512 Gaussian components, the rank of *i-vector* is fixed at 100 and the LDA dimension is fixed at 10.

IV. Experimental Results and Discussion

The performance of dysarthric level assessment systems is measured using detection accuracy rate (DAR) and also analyzed using confusion matrices.

A. Performance of the Baseline System

In the MFCC feature extraction process [46], the triangular filters are placed in a nonlinear scale to map the frequency bins in Hz to Mel-scale following the human speech perception. The Mel-scale warping emphasizes the lower frequency bins than the higher frequency bins [46]. Consequently, the fine level features those lie in the higher frequency range may not be captured by the MFCC feature. Alternatively, the linearly spaced filterbank provides equal emphasis to all the frequency bins [52], [65]. To study the impact of Mel and linearly spaced filterbank on the cepstral coefficients for the task of dysarthric level assessment, we have extracted the cepstral coefficients by replacing the Mel-frequency triangular filterbank with linearly spaced triangular filterbank in the feature extraction process. The cepstral coefficients extracted using linear filterbank are termed as linear frequency cepstral coefficient (LFCC) [46], [66].

The performances of the dysarthric level assessment system for MFCC and LFCC features on the SD_Test and SI_Test datasets are summarized in Table II. In this study, MFCC and LFCC features are extracted using 40 channel filterbank [28]. For both kinds of features, the DAR observed for SI_Test is less compared to the SD_Test. As mentioned earlier, the *i-vectors* captures the speaker factors along with the information of dysarthria. Since the speakers present in Train_Data and SD_Test are the same, the speaker factor present in the *i-vectors* is normalized up to a great extent. From these preliminary experimental results, it is evident that the cosine kernel-based scoring provides better DAR than the PLDA-based scoring for the SI_Test dataset. The WCCN followed by LDA provides improved DAR than WCCN and LDA-based projection. For both the test datasets, the performance observed for the MFCC feature is better than the LFCC. Therefore, further studies are performed using the Mel-frequency filterbank.

TABLE II. The Performance of the Dysarthric Assessment System Using MFCC and LFCC Features Extracted Using 40 Channel Filterbank. The Performance Is Given in Terms of DAR (in %) for Cosine Kernel and PLDA Based Scoring Schemes

Test Dataset	Feature Type	Cosine Kernel			PLDA
		LDA	WCCN	LDA-WCCN	LDA
SD_Test	MFCC	87.529	88.280	91.436	91.747
	LFCC	86.933	87.123	88.883	89.209
SI_Test	MFCC	46.562	47.719	48.886	47.104
	LFCC	43.948	45.089	46.969	45.854

1. Impact of Mel-Filterbank Size on Dysarthria Discrimination of MFCC Feature

As discussed earlier, the information about dysarthria is present at a fine level in the short-term magnitude spectra. Consequently, the dysarthria discrimination of the MFCC feature may be enhanced up to an extent by increasing the size of the Mel-filterbank during the feature computation process. The performance of the dysarthric level assessment system employing the MFCC feature for the varied size of Mel-filterbank is given in Table III. It is evident that the DAR for the SI_Test dataset employing the MFCC feature improves with an increase in the size of the Mel-filterbank and the best DAR is observed when the Mel-filterbank size is 160. On further improving the size of the filterbank, it captures the redundancy present in the magnitude spectra. On the other hand, by increasing the filterbank size, DAR reduces for the SD_Test dataset. *Most of the practical applications demand assessment of dysarthric levels in speaker and text-independent mode.* Therefore, this study emphasizes the SI_Test mode of operation. Further studies on the MFCC feature is presented using 160 channel Mel-filterbank.

TABLE III. The Performance of the Dysarthric Assessment System Employing MFCC Feature for Different Sizes of Mel- Filterbank. The Performance Is Given in Terms of DAR (in %) for Cosine Kernel and PLDA Based Scoring Schemes

Filterbank Size	SD_Test		SI_Test	
	Cosine Kernel	PLDA	Cosine Kernel	PLDA
	LDA-WCCN	LDA	LDA-WCCN	LDA
40	91.436	91.747	48.886	47.104
80	90.247	90.934	52.448	50.489
120	89.622	90.121	54.500	51.875
160	88.090	88.933	56.646	54.563
200	88.895	89.214	55.812	52.677
240	89.064	89.838	53.636	51.864

2. Impact of Cepstral Liftering on Dysarthria Discrimination of MFCC Feature

In the MFCC feature extraction process, cepstral liftering is used to estimate the spectral envelope that represents the resonance structure of the vocal tract system [46], [53]. On the other hand, the cepstral liftering operation smooths out the pitch harmonics [67], [68]. The pitch harmonics may contain information about dysarthria. To capture the effect of pitch harmonics in the MFCC feature, instead of using fixed 13 dimensional base cepstral coefficients, we have studied the performance of the dysarthric level assessment system employing different sizes of base MFCC. In this study, the MFCC feature is extracted using 160 channel Mel-filterbank. The performance of the dysarthric level assessment system for varied dimensions of base MFCC feature is given in Table IV. For both cosine kernel and PLDA based scoring schemes, on increasing the dimension of cepstral coefficients the performance of the dysarthric level assessment system is improved for the SD_Test dataset. On the other hand, for the SI_Test dataset, the best performance is observed for 13 dimensional base MFCC feature. It may be due to modeling speaker information instead of dysarthria. These experimental results show that the optimal performance for the SI_Test dataset is observed for the 13 dimensional base MFCC feature extracted using 160 channel Mel-filterbank.

TABLE IV. The Performance of the Dysarthric Assessment System Employing Varied Dimensions of MFCC Feature Extracted Using 160 Mel-Filterbank. The Performance is Given in Terms of DAR (in %) for Cosine Kernel and PLDA Based Scoring Schemes

Feature Dim.	SD_Test		SI_Test	
	Cosine Kernel	PLDA	Cosine Kernel	PLDA
	LDA-WCCN	LDA	LDA-WCCN	LDA
8	84.995	85.106	51.625	50.048
13	88.090	88.933	56.646	54.563
18	89.967	90.529	48.614	44.761
23	90.935	91.841	44.083	41.281

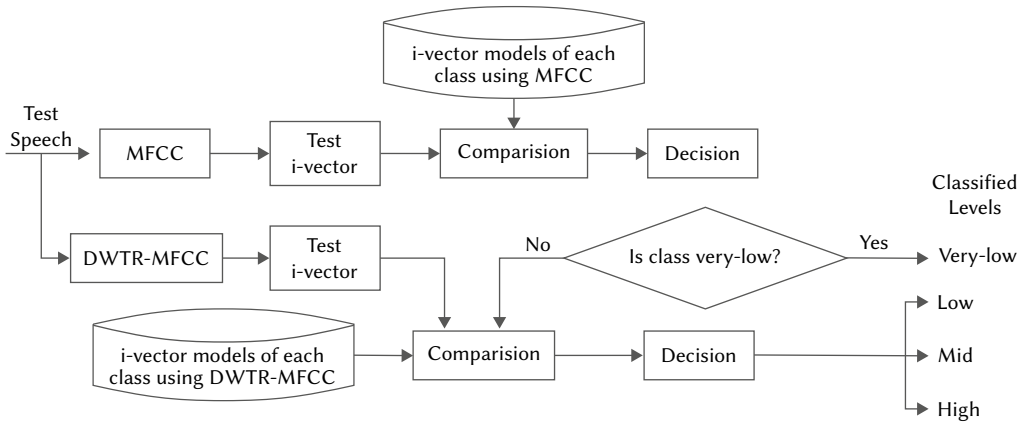


Fig. 4. Block diagram illustrating the proposed scheme for combining the efficacy of the MFCC and DWTR-MFCC feature to improve the performance of the dysarthric level assessment system.

B. Performance of Dysarthric Level Assessment System Using DWTR-MFCC Feature

The performance of the developed dysarthric level assessment system employing the DTWR-MFCC feature is given in Table V. For performance comparison, the DAR obtained for the MFCC with optimal feature parameter selection, LPCC [53], CQCC [37], SMAC [54] features is also given in the Table. 5. Similar to the MFCC, the explored features also provide improved performance for the cosine kernel-based scoring than the PLDA-based scoring scheme. Furthermore, for all the explored features, a reduced DAR is observed for the SI_Test dataset compared to the SD_Test dataset. In the case of the SI_Test dataset, CQCC provides improve performance than LPCC and SMAC features. On the other hand, for the SD_Test dataset, SMAC provides better performance than LPCC and CQCC features. However, the MFCC feature extracted using optimal parameter selection provides better performance than all the explored features for the SI_Test dataset. In the case of the SI_Test dataset, the usage of the proposed DTWR-MFCC feature improves performance compared to the MFCC feature. As discussed earlier, in most of the practical applications, the dysarthric level assessment needs to be done in speaker and text-independent mode. Therefore, the proposed way of computing the cepstral coefficient is more effective than the conventional MFCC feature.

TABLE V. The Performance of the Dysarthric Assessment System is Given for Proposed DWTR-MFCC and Explored Features. The Performance is Given in Terms of DAR (in %) for Cosine Kernel and PLDA Based Scoring Schemes

Filterbank Size	SD_Test		SI_Test	
	Cosine Kernel	PLDA	Cosine Kernel	PLDA
	LDA-WCCN	LDA	LDA-WCCN	LDA
DWTR-MFCC	85.620	86.589	60.094	57.271
MFCC	88.090	88.933	56.646	54.563
LPCC	92.904	93.436	50.323	48.531
CQCC	88.367	88.972	53.989	49.875
SMAC	93.823	94.017	47.458	44.271

For further analysis, we have computed the confusion matrices obtained using the DWTR-MFCC and MFCC features. The confusion matrices obtained for the MFCC and DWTR-MFCC features on the SI_Test dataset are given in Table VI. Table VI (a) and Table VI (b) represents the confusion matrices obtained for cosine kernel scoring on LDA-WCCN projected *i-vectors* for the MFCC and DWTR-MFCC, respectively. Table VI (c) and Table VI (d) represents the PLDA scoring on LDA projected *i-vectors* for MFCC and DWTR-MFCC, respectively.

By comparing the confusion matrices obtained by the DWTR-MFCC feature with the MFCC features given in Table VI, it can be observed that the DWTR-MFCC feature is more discriminating for mid and low dysarthric classes. On the other hand, the MFCC feature is more discriminating for the very-low dysarthric level. Motivated by this observation, next we have studied the possibilities of improving the overall performance of the developed dysarthric level assessment system by employing discriminating power of both kinds of features.

TABLE VI. Confusion Matrices are Given for the *I-vector* Based Dysarthric Level Assessment Systems Using the Proposed DWTR-MFCC Feature. This Analysis is Given on the SI_Test Dataset. (A) and (B) Cosine Kernel Scoring on LDA-WCCN Projected *I-Vectors* for MFCC and DWTR-MFCC, Respectively. (C) and (D) PLDA Scoring on LDA Projected *I-Vectors* for MFCC and DWTR-MFCC, Respectively. The Abbreviation H, M, L, and VL Refer to the High, Mid, Low, and Very-Low Speech Intelligibility Groups, Respectively

(a)

True class \ Predicted class	H	M	L	VL
H	83.083	10.084	4.500	2.333
M	23.750	50.500	19.500	6.250
L	7.750	43.500	39.500	9.250
VL	20.750	8.875	16.875	53.500

DAR: 56.646 %

(c)

True class \ Predicted class	H	M	L	VL
H	75.500	18.500	4.333	1.667
M	11.250	68.750	16.500	3.500
L	2.750	59.750	34.500	3.000
VL	14.250	26.250	20.000	39.500

DAR: 54.563 %

(b)

True class \ Predicted class	H	M	L	VL
H	83.750	9.833	3.250	3.167
M	18.500	58.250	12.250	11.000
L	5.500	26.250	61.750	6.500
VL	23.750	4.750	34.875	36.625

DAR: 60.094 %

(d)

True class \ Predicted class	H	M	L	VL
H	70.833	18.083	7.167	3.917
M	7.750	68.250	16.250	7.750
L	2.000	34.250	60.000	3.750
VL	13.000	10.500	46.500	30.000

DAR: 57.271 %

C. Performance of Dysarthric Level Assessment Systems Using Proposed Two-Stage Classification Scheme

To utilize the classification efficacy of the MFCC and DWTR-MFCC features, in this study at the first level, the classification scores obtained for both kinds of features are combined prior to decision. Next, a two-stage classification scheme is proposed to fully utilize the classification efficacy of both kinds of features. The block diagram representation of the proposed two-stage classification scheme is shown in Fig. 4. As shown in the figure, at the first stage a given test utterance is classified employing the MFCC feature. Depending on the

maximum score, if the assigned class is a very-low intelligible group then the test data is assigned to that class and the classification process is terminated. On the other hand, if the assigned class is any other intelligible group (high, mid, and low), further classification among these groups is performed employing the dysarthric level assessment system developed using the DWTR-MFCC feature, and the final class assignment is done depending on the maximum score obtained for test data.

TABLE VII. The Performance Dysarthric Level Assessment System is Given for Score Level Combination and the Proposed Combination Scheme. The Performance is Given in Terms of DAR (in %) for Cosine Kernel and PLDA Based Scoring Schemes

Feature Type	SD_Test		SI_Test	
	Cosine Kernel LDA-WCCN	PLDA LDA	Cosine Kernel LDA-WCCN	PLDA LDA
MFCC	88.090	88.933	56.646	54.563
DWTR-MFCC	85.620	86.589	60.094	57.271
Score-comb	89.560	90.465	61.261	58.823
Proposed	87.122	88.026	65.813	60.750

TABLE VIII. Confusion Matrices are Given for the *I-Vector* Based Dysarthric Level Assessment Systems Using Score Combination and Proposed Approach, Respectively. This Analysis is Given on the SI_Test Dataset. (A) and (B) Cosine Kernel Scoring on LDA-WCCN Projected *I-Vectors* for Score Combination and Proposed Method, Respectively. (C) and (D) PLDA Scoring on LDA Projected I-Vectors for Score Combination and Proposed Method, Respectively. The Abbreviation H, M, L, and VL Refer to the High, Mid, Low, and Very-Low Speech Intelligibility Groups, Respectively

(a)

True class \ Predicted class	H	M	L	VL
H	88.417	6.833	2.750	2.000
M	19.750	57.000	15.750	7.500
L	5.000	34.500	52.250	8.250
VL	22.375	6.250	24.000	47.375

DAR: 61.261 %

(c)

True class \ Predicted class	H	M	L	VL
H	76.917	16.167	5.083	1.833
M	7.500	75.000	12.750	4.750
L	2.250	47.250	46.000	4.500
VL	14.125	18.750	29.750	37.375

DAR: 58.823 %

(b)

True class \ Predicted class	H	M	L	VL
H	86.000	8.167	3.500	2.333
M	16.500	63.000	14.250	6.250
L	4.250	25.750	60.750	9.250
VL	18.625	3.250	24.625	53.500

DAR: 65.813%

(d)

True class \ Predicted class	H	M	L	VL
H	73.250	17.250	7.833	1.667
M	7.750	70.250	18.500	3.500
L	2.000	35.000	60.000	3.000
VL	11.000	9.250	40.250	39.500

DAR: 60.750%

The DAR obtained using the score level combination and proposed two-stage classification scheme is summarized in Table VII. The score level combination improves the DAR compared to individual features for both cosine kernel and PLDA based scoring methods. The DAR is further improved on the use of the proposed two-stage classification scheme. For cosine kernel-based scoring on LDA-WCCN projected *i-vectors*, the DAR improved to 65.813% in case of the SI_Test dataset. To further analyze the merits of the proposed two-stage classification scheme, the confusion matrices obtained for the proposed classification scheme are compared with the score level combination. Table VIII (a) and Table VIII (b) represents the confusion matrices obtained for cosine kernel scoring on LDA-WCCN projected *i-vectors* for score level combination and proposed approach, respectively. Table VIII (c) and Table VIII (d) represents the PLDA scoring on LDA projected *i-vectors* for score level combination and proposed method, respectively. By comparing the confusion matrices obtained for both approaches, it can be observed that in case of the proposed two-stage classification method the classification accuracy of each class is improved. This is mainly due to the complete utilization of the discrimination power of both kinds of the feature. This experimental result also shows that the proposed DWTR-MFCC carries additional information than MFCC.

V. Conclusion

The work presented in this paper aims at improving the performance of an automatic dysarthric level assessment system by capturing the fine level information present in different sub-bands of the speech signal. To capture the fine level information, firstly, the performance of the dysarthric level assessment system employing the MFCC feature is studied by varying the shape and size of the triangular filterbank. Next, for a better representation of sub-band information, the input speech signal is decomposed into four levels using DWT decomposition. For each input speech signal, five speech signals representing different sub-bands are then reconstructed using IDWT. The log filterbank energies are computed by analyzing the DFT magnitude spectra of each reconstructed speech using a 30-channel Mel-filterbank. For each analysis frame, the log filterbank energies obtained across all reconstructed speech are combined, and DCT is performed to represent the cepstral feature, termed as DWTR-MFCC in this study. The performance of *i-vector* based four-level dysarthric assessment system on the UA-Speech corpus shows that the overall performance of the system employing the MFCC feature improves by increasing the size of Mel-filterbank. However, a large overlapping between mid and low dysarthric levels is observed. On the use of the DWTR-MFCC feature, performance of the developed dysarthric level assessment is further improved by reducing the overlapping between mid and low dysarthric levels. But, reduced classification accuracy is observed for very-low dysarthric levels due to miss classification of very-low dysarthric level to low dysarthric level. Motivated by these observations, finally, a two-stage classification approach is proposed by employing the efficacy of MFCC and DWTR-MFCC features. The proposed approach improves the classification accuracy of the developed dysarthric level assessment system by reducing the overlapping between any two classes without loss of performance for individual features (MFCC or DWTR-MFCC).

References

[1] J. R. Duffy, *Motor speech disorders e-book: Substrates, differential diagnosis, and management*. Elsevier Health Sciences, 2019.

[2] R. Sandyk, "Resolution of dysarthria in multiple sclerosis by treatment with weak electromagnetic fields," *International Journal of Neuroscience*, vol. 83, no. 1-2, pp. 81–92, 1995.

[3] J. Müller, G. K. Wenning, M. Verny, A. McKee, K. R. Chaudhuri, K. Jellinger, W. Poewe, I. Litvan, "Progression of dysarthria and dysphagia in postmortem-confirmed parkinsonian disorders," *Archives of neurology*, vol. 58, no. 2, pp. 259–264, 2001.

[4] S. Skodda, H. Rinsche, U. Schlegel, "Progression of dysprosody in Parkinson's disease over time—a longitudinal study," *Movement disorders: official journal of the Movement Disorder Society*, vol. 24, no. 5, pp. 716–722, 2009.

[5] J. B. Polikoff, H. T. Bunnell, "The nemours database of dysarthric speech: A perceptual analysis," in *Proc. ICPS*, 1999, pp. 783–786.

[6] R. D. Kent, "Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders," *American Journal of Speech-Language Pathology*, vol. 5, no. 3, pp. 7–23, 1996.

[7] G. Constantinescu, D. Theodoros, T. Russell, E. Ward, S. Wilson, R. Wootton, "Assessing disordered speech and voice in Parkinson's disease: a telerehabilitation application," *International journal of language & communication disorders*, vol. 45, no. 6, pp. 630–644, 2010.

[8] K. K. Baker, L. O. Ramig, E. S. Luschei, M. E. Smith, "Thyroarytenoid muscle activity associated with hypophonia in parkinson disease and

aging," *Neurology*, vol. 51, no. 6, pp. 1592–1598, 1998.
[9] S. Skodda, W. Visser, U. Schlegel, "Vowel articulation in Parkinson's disease," *Journal of Voice*, vol. 25, no. 4, pp. 467–472, 2011.
[10] A. Maier, T. Haderlein, F. Stelzle, E. Nöth, E. Nkenke, F. Rosanowski, A. Schützenberger, M. Schuster, "Automatic speech recognition systems for the evaluation of voice and speech disorders in head and neck cancer,"*EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, pp. 1–7, 2009.
[11] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, E. Nöth, "Peaks–a system for the automatic evaluation of voice and speech disorders," *Speech Communication*, vol. 51, no. 5, pp. 425–437, 2009.
[12] K. Riedhammer, G. Stemmer, T. Haderlein, M. Schuster, F. Rosanowski, E. Noth, A. Maier, "Towards robust automatic evaluation of pathologic telephone speech," in *Proc. ASRU (Workshop)*, 2007, pp. 717–722.
[13] M. J. Kim, Y. Kim, H. Kim, "Automatic intelligibility assessment of dysarthric speech using phonologically- structured sparse linear model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 694–704, 2015.
[14] C. Deng, G. Lai, H. Deng, "Improving word vector model with part-of-speech and dependency grammar information," *CAAI Transactions on Intelligence Technology*, vol. 5, no. 4, pp. 276–282, 2020.
[15] F. Rudzicz, "Comparing speaker-dependent and speaker-adaptive acoustic models for recognizing dysarthric speech," in *Proc. International ACM SIGACCESS on Computers and Accessibility*, 2007, pp. 255–256.
[16] F. Rudzicz, "Phonological features in discriminative classification of dysarthric speech," in *Proc. ICASSP*, 2009, pp. 4605–4608.
[17] K. L. Kadi, S. A. Selouani, B. Boudraa, M. Boudraa, "Automated diagnosis and assessment of dysarthric speech using relevant prosodic features," in *Transactions on Engineering Technologies*, 2014, pp. 529– 542.
[18] C. Bhat, B. Vachhani, S. K. Kopparapu, "Automatic assessment of dysarthria severity level using audio descriptors," in *Proc. ICASSP*, 2017, pp. 5070–5074.
[19] M. Perez, W. Jin, D. Le, N. Carlozzi, P. Dayalu, A. Roberts, E. M. Provost, "Classification of huntington disease using acoustic and lexical features," in *Proc. INTERSPEECH*, 2018, pp. 1898–1902.
[20] N. Saleem, M. I. Khattak, "Deep neural networks for speech enhancement in complex-noisy environments," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 1, pp. 84–90, 2020.
[21] N. Saleem, M. I. Khattak, E. Verdú, "On improvement of speech intelligibility and quality: A survey of unsupervised single channel speech enhancement algorithms," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 2, pp. 78–89, 2020.
[22] B. Yang, R. Ding, Y. Ban, X. Li, H. Liu, "Enhancing direct-path relative transfer function using deep neural network for robust sound source localization," *CAAI Transactions on Intelligence Technology*, pp. 1–9, 2021.
[23] M. J. Kim, B. Cao, K. An, J. Wang, "Dysarthric speech recognition using convolutional LSTM neural network," in *Proc. INTERSPEECH*, 2018, pp. 2948–2952.
[24] H. Liu, P. Yuan, B. Yang, G. Yang, Y. Chen, "Head-related transfer function–reserved time- frequency masking for robust binaural sound source localization," *CAAI Transactions on Intelligence Technology*, pp. 26–33, 2021.
[25] A. A. Joshy, R. Rajan, "Automated dysarthria severity classification using deep learning frameworks," in *Proc. European Signal Processing Conference*, 2021, pp. 116–120.
[26] C. Bhat, H. Strik, "Automatic assessment of sentence- level dysarthria intelligibility using BLSTM," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 322–330, 2020.
[27] K. Gurugubelli, A. K. Vuppala, "Perceptually enhanced single frequency filtering for dysarthric speech detection and intelligibility assessment," in *Proc. ICASSP*, 2019, pp. 6410–6414.
[28] K. Gurugubelli, A. K. Vuppala, "Analytic phase features for dysarthric speech detection and intelligibility assessment," *Speech Communication*, vol. 121, pp. 1–15, 2020.
[29] S.-A. Selouani, H. Dahmani, R. Amami, H. Hamam, "Using speech rhythm knowledge to improve dysarthric speech recognition," *International Journal of Speech Technology*, vol. 15, no. 1, pp. 57–64, 2012.
[30] J. Kim, N. Kumar, A. Tsiartas, M. Li, S. S. Narayanan, "Automatic intelligibility classification of sentence- level pathological speech," *Computer Speech & Language*, vol. 29, no. 1, pp. 132–144, 2015.
[31] K. Kadi, S. A. Selouani, B. Boudraa, M. Boudraa, "Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge," *Biocybernetics and Biomedical Engineering*, vol. 36, no. 1, pp. 233–247, 2016.
[32] A. Benba, A. Jilbab, A. Hammouch, "Detecting patients with Parkinson's disease using Mel frequency cepstral coefficients and support vector machines," *International Journal on Electrical Engineering and Informatics*, vol. 7, no. 2, pp. 297–307, 2015.
[33] H. Chandrashekar, V. Karjigi, N. Sreedevi, "Investigation of different time-frequency representations for intelligibility assessment of dysarthric speech," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 12, pp. 2880–2889, 2020.
[34] D. Martínez, E. Lleida, P. Green, H. Christensen, A. Ortega, A. Miguel, "Intelligibility assessment and speech recognizer word accuracy rate prediction for dysarthric speakers in a factor analysis subspace," *ACM Transactions on Accessible Computing*, vol. 6, no. 3, pp. 1–21, 2015.
[35] A. Benba, A. Jilbab, A. Hammouch, "Discriminating between patients with Parkinson's and neurological diseases using cepstral analysis," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 10, pp. 1100–1108, 2016.
[36] S. Oue, R. Marxer, F. Rudzicz, "Automatic dysfluency detection in dysarthric speech using deep belief networks," in *Proc. SLPAT*, 2015, pp. 60–64.
[37] J. C. Brown, "Calculation of a constant Q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.
[38] M. Little, P. McSharry, E. Hunter, J. Spielman, L. Ramig, "Suitability of dysphonia measurements for telemonitoring of parkinson's disease," *Nature Precedings*, pp. 1–27, 2008.
[39] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, pp. 1264–1271, 2012.
[40] I. Kodrasi, H. Bourlard, "Super-gaussianity of speech spectral coefficients as a potential biomarker for dysarthric speech detection," in *Proc. ICASSP*, 2019, pp. 6400–6404.
[41] N. Narendra, P. Alku, "Dysarthric speech classification using glottal features computed from non-words, words and sentences," in *Proc. INTERSPEECH*, 2018, pp. 3403–3407.
[42] N. Narendra, P. Alku, "Dysarthric speech classification from coded telephone speech using glottal features," *Speech Communication*, vol. 110, pp. 47–55, 2019.
[43] N. Narendra, P. Alku, "Automatic assessment of intelligibility in speakers with dysarthria from coded telephone speech using glottal features," *Computer Speech & Language*, vol. 65, pp. 1–14, 2021.
[44] S. G. Mallat, "A theory for multiresolution signal decomposition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 674–693, 1989.
[45] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," *IEEE Transactions on information theory*, vol. 36, no. 5, pp. 961–1005, 1990.
[46] S. Davis, P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.
[47] P. Singh, G. Pradhan, S. Shahnawazuddin, "Denoising of ecg signal by non-local estimation of approximation coefficients in dwt," *Biocybernetics and Biomedical Engineering*, vol. 37, no. 3, pp. 599–610, 2017.
[48] H. Ayad, M. Khalil, "Qam-dwt-svd based watermarking scheme for medical images," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 3, pp. 81–89, 2018.
[49] R. Abbasi, M. Esmaeilpour, "Selecting statistical characteristics of brain signals to detect epileptic seizures using discrete wavelet transform and perceptron neural network," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 5, pp. 33–38, 2017.
[50] A. K. H. Al-Ali, D. Dean, B. Senadji, V. Chandran, G. R. Naik, "Enhanced forensic speaker verification using a combination of dwt and mfcc feature warping in the presence of noise and reverberation conditions," *IEEE Access*, vol. 5, pp. 15400–15413, 2017.
[51] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang,

K. Watkin, S. Frame, "Dysarthric speech database for universal access research," in *Proc. INTERSPEECH*, 2008, pp. 1741– 1744.

[52] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy- Wilson, S. Shamma, "Linear versus mel frequency cepstral coefficients for speaker recognition," in *Proc. ASRU (Workshop)*, 2011, pp. 559–564.

[53] L. R. Rabiner, R. W. Schafer, "Introduction to digital speech processing," *Foundations and Trends in Signal Processing*, vol. 1, no. 1–2, pp. 1–194, 2007.

[54] P. Tsiakoulis, A. Potamianos, D. Dimitriadis, "Spectral moment features augmented by low order cepstral coefficients for robust asr," *IEEE Signal Processing Letters*, vol. 17, no. 6, pp. 551–554, 2010.

[55] K. Maity, G. Pradhan, J. P. Singh, "A pitch and noise robust keyword spotting system using smac features with prosody modification," *Circuits, Systems, and Signal Processing*, vol. 40, no. 4, pp. 1892–1904, 2021.

[56] J. G. Wilpon, L. R. Rabiner, T. Martin, "An improved word-detection algorithm for telephone-quality speech incorporating both syntactic and semantic constraints," *AT&T Bell Laboratories Technical Journal*, vol. 63, no. 3, pp. 479–498, 1984.

[57] O. Viikki, K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1-3, pp. 133–147, 1998.

[58] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[59] D. Garcia-Romero, C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. INTERSPEECH*, 2011, pp. 249–252.

[60] G. Pradhan, S. R. M. Prasanna, "Speaker verification by vowel and nonvowel like segmentation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 854–867, 2013.

[61] S. Balakrishnama, A. Ganapathiraju, "Linear discriminant analysis-a brief tutorial," *Institute for Signal and information Processing*, vol. 18, pp. 1–8, 1998.

[62] A. O. Hatch, S. S. Kajarekar, A. Stolcke, "Within- class covariance normalization for svm-based speaker recognition.," in *Proc. INTERSPEECH*, 2006, pp. 1471– 1474.

[63] N. Dehak, R. Dehak, J. R. Glass, D. A. Reynolds, P. Kenny, *et al.*, "Cosine similarity scoring without score normalization techniques.," in *Proc. Odyssey*, 2010, pp. 1–5.

[64] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, P. Dumouchel, "Plda for speaker verification with utterances of arbitrary duration," in *Proc. ICASSP*, 2013, pp. 7649–7653.

[65] Y. Jung, Y. Kim, H. Lim, H. Kim, "Linear-scale filterbank for deep neural network-based voice activity detection," in *Proc. O-COCOSDA*, 2017, pp. 1–5.

[66] S. Debnath, P. Roy, "Audio-visual automatic speech recognition using pzm, mfcc and statistical analysis," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 2, pp. 121–133, 2021.

[67] B. Pattanayak, G. Pradhan, "Pitch-robust acoustic feature using single frequency filtering for children's kws," *Pattern Recognition Letters*, vol. 150, pp. 183–188, 2021.

[68] J. G. Wilpon, C. N. Jacobsen, "A study of speech recognition for children and the elderly," in *Proc. ICASSP*, vol. 1, 1996, pp. 349–352.

Laxmi Priya Sahu

Laxmi Priya Sahu received her B.Tech. degree in Electronics and Telecommunication Engineering from Biju Patnaik University of Technology (BPUT), Odisha, India, in 2012 and M.Tech. degree in Communication System Engineering from KIIT University, Odisha, India, in 2015. She is pursuing Ph.D. in the Department of Electronics and Communication Engineering at National Institute of Technology Patna, India. Her research interests are speech signal analysis, and biomedical signal processing.

Gayadhar Pradhan

Gayadhar Pradhan received his M.Tech. and Ph.D. degrees in Electronics and Electrical Engineering from Indian Institute of Technology Guwahati, India, in 2009 and 2013, respectively. He is currently working as Associate professor in the Department of Electronics and Communication Engineering at National Institute of Technology Patna, India. His research interests are speech signal processing, speaker recognition and speech recognition.

Jyoti Prakash Singh

Jyoti Prakash Singh is an assistant professor and the Department Head of Computer Science and Engineering at the National Institute of Technology Patna in India. He has co-authored seven textbooks and one edited book with McGraw Hill, Elsevier, and Springer, among others. He has over 45 international journal publications in leading publishers, as well as over 50 international conference proceedings. He was a co-investigator in an MietY-sponsored project to develop algorithms for spam/fake calls in telephone conversations. His research interests include social media mining, deep learning, data security, and speech processing. He is an Associate Editor for the International Journal of Electronic Government Research. In 2020, he received the S4DS Data Scientist (Academia) Award from the Society for Data Science.

A Fuzzy-Based Multimedia Content Retrieval Method Using Mood Tags and Their Synonyms in Social Networks

Chang Bae Moon[1], Jong Yeol Lee[2], Byeong Man Kim[2] *

[1] ICT-Convergence Research Center, Kumoh National Institute of Technology (Korea)
[2] Computer Software Engineering, Kumoh National Institute of Technology (Korea)

Received 4 March 2021 | Accepted 12 September 2021 | Published 3 October 2022

unir
LA UNIVERSIDAD EN INTERNET

Abstract

The preferences of Web information purchasers are rapidly evolving. Cost-effectiveness is now becoming less regarded than cost-satisfaction, which emphasizes the purchaser's psychological satisfaction. One method to improve a user's cost-satisfaction in multimedia content retrieval is to utilize the mood inherent in multimedia items. An example of applications using this method is SNS (Social Network Services), which is based on folksonomy, but its applications encounter problems due to synonyms. In order to solve the problem of synonyms in our previous study, the mood of multimedia content is represented with arousal and valence (AV) in Thayer's two-dimensional model as its internal tag. Although some problems of synonyms could now be solved, the retrieval performance of the previous study was less than that of a keyword-based method. In this paper, a new method that can solve the synonym problem is proposed, while simultaneously maintaining the same performance as the keyword-based approach. In the proposed method, a mood of multimedia content is represented with a fuzzy set of 12 moods of the Thayer model. For the analysis, the proposed method is compared with two methods, one based on AV value and the other based on keyword. The analysis results demonstrate that the proposed method is superior to the two methods.

Keywords

Cost-Satisfaction, Mood Fuzzy Value, Multimedia Content Mood Tag, Multimedia Content Retrieval, Social Network.

DOI: 10.9781/ijimai.2022.10.005

I. Introduction

ACCORDING to [1], consumption patterns are changing from cost-effectiveness, which refers to performance with respect to price, to cost-satisfaction, which emphasizes cost-effectiveness along with psychological satisfaction. The multimedia content search method needs to be changed according to this flow. It will not only increase psychological satisfaction through searching by mood or emotion rather than text content, but also provide various additional services. Several studies have been conducted in accordance with this trend. A set of works [2]-[4] define the mood model, another research [5] shows how to search for multimedia content using the mood, and other papers [6], [7] analyze the relationship between mood and color of multimedia content.

Psychological satisfaction is a new trend in modern society as the single-person society has developed recently, such as the Yolo [8] and the Solo [9], which means that consumer's consumption prioritizes consumer's satisfaction over content's price. That is, cost-effectiveness means price satisfaction on multimedia content and the psychological satisfaction means customer satisfaction on its service. In this context, we propose a multimedia content retrieval method using mood to increase user's psychological satisfaction.

There are four main types of multimedia content search methods applicable to SNS (Social Network Services). The first one is query-by-text that retrieves information matching the query based on textual information stored in the multimedia content database. The second one is query-by-part where a part of the content (for example, part of music) is served as a query and content similar to that part is found. The third one is the query-by-example, which accepts the content itself as a query to find similar content. Unlike query-by-part, the entire content is entered as a query rather than a part. The last fourth one is query-by-class, which retrieves the content of a given class as input using predefined genres or classes.

Among the four types, query-by-class is a common search method, but requires expert or curator participation. That is, when new multimedia content is provided, class information must be input. In a situation where new multimedia content is constantly and rapidly generated, this method is difficult to respond quickly. A possible solution is to categorize contents using tags according to the taxonomy, enter bibliographic information about the class, or assign a class code. However, this method still requires human curator intervention. In addition, if the knowledge of a specific topic is insufficient, the classification system cannot be expanded, and one or more classes can be assigned to some content, so separate processing is required.

The folksonomy is mentioned as an alternative to the taxonomy-based classification system. It has a flat structure and requires users to participate in lieu of professional management of a librarian or curator [10]. Famous SNS like Instagram, Facebook, Last.fm, IPTV and YouTube use this classification to provide services. The biggest advantage of the folksonomy is that it avoids expandability and monopoly, which is a problem with the taxonomy-based classification systems. However, the folksonomy has a synonymous problem between tags used to describe content. That is, it is necessary to identify similar tags as well

* Corresponding author.
E-mail addresses: cb.moon@kumoh.ac.kr (C. B. Moon), soyeum@kumoh.ac.kr (J. Y. Lee), bmkim@kumoh.ac.kr (B. M. Kim).

as content including query tags and recommend content with similar tags. Also, if you provide a new word as a query tag, the content is not searched. In this case, you need to identify new words and recommend related content using existing tags associated with the new words. To solve this, a method that considers analogues based on AV is proposed in the study of [10]. Unfortunately, this method achieves performance that is inferior to a method based on keywords at recall level 0.1. Users use search engines to get the information they want, and they usually tend to see only the first few pages of search results. Therefore, it is important to provide a high-accuracy search at low recall levels. In this context, this paper attempts to improve the precision at low recall levels in [10].

In this paper, a method is proposed that supplements the problem of SNS service and can approach the retrieval performance of the keyword-based approach at recall level 0.1. To achieve this, the mood of a multimedia content is represented with a fuzzy set of 12 moods of the Thayer model. Then, the association between the fuzzy vectors of contents and the fuzzy vectors of tags of contents is analyzed. In addition, for the performance comparison, the proposed method is compared with two methods, one based on AV value [10] and the other based on keyword (Last.fm approach). Additional experiments are also carried out for four different similarity measures between two fuzzy vectors [11], [12], [13], [14]. Although the proposed method in this paper is applicable to all types of multimedia content if they only include annotation tags, it is preferentially applied to music content like the previous study for performance analysis.

II. Relevant Works

Among the methods used to define the mood, there are Russell [2] and Hevner [3]. These methods may cause semantically overlapping or ambiguous problems due to the adjective expression. A model to solve this ambiguity problem is the model of Thayer [4]. In this model, 12 mood words are expressed as two-dimensional vectors of Arousal and Valence, and the mood of multimedia content is expressed using these mood vectors. Arousal represents the intensity of stimuli felt by multimedia consumers, and Valence represents the stability experienced by multimedia consumers (Fig. 1) [10], [15]. Among the multimedia content retrieval methods, some [5]-[7] used this model.

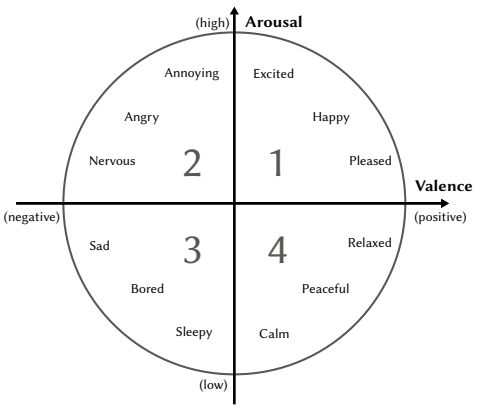


Fig. 1. Thayer's two-dimensional model.

By use of Thayer's mood vector (i.e., AV vector), Moon et al. [5] tried to solve the problems of music folksonomy-problems of tagging level (for example, happiness, very happy, too happy, etc.), synonym problems, and new words. In this study, in order to express music as AV vector values, subjects were asked to enter the Arousal and Valence values for the music they heard after listening to the music. So, both music and tags were expressed as AV values, so music with synonym tags could be searched. That is, the problem of synonyms in folksonomy could be solved naturally. However, this approach was difficult to apply to large amounts of data because the subjects had to assign a two-dimensional mood vector to each piece of music.

Other works [16]-[18] used SVM (Support Vector Machine), one of the machine learning methods, to grasp the mood of a new music piece. The learning data is composed of inputs and the corresponding output values. In these studies, the acoustic and musical characteristics of a music piece are used as input values, and the mood values are used as output values. For the mood values of music pieces, folksonomy mood tags provided by Last.fm were used. Some authors [19]-[21] studied the automatic tagging method for web content. In order to easily input the tags of web contents, tags related to web contents are presented and related tags can be selected among them. Others [22]–[27] conducted an automatic tagging study on images. After learning the association between the properties included in the content and the image tags using SVM, Bayes Point Machines [25], etc., the tags for the new image were automatically added using this.

Among existing AV-based retrieval methods to solve the problem of synonyms, there is one that [10] obtains mood vectors (strength of arousal and valence) of music from social and international mood data which were built for several years, and defines AV values of tags in order to make synonym-based search possible. The study, however, reveals that retrieval performance is worse than that of a keyword-based approach at recall level 0.1. In this paper, we propose a retrieval method that uses fuzzy-based multimedia mood which achieves the same retrieval performance as a keyword-based approach at recall level 0.1. In addition, the excellence of the proposed method is demonstrated as it was comparatively analyzed with the AV-based retrieval method of [10] and a keyword-based approach.

There are some works [28], [29], [30] on detecting mood based on fuzzy theory. In a research [28], the music mood detection using the AV model based on a fuzzy theory is proposed, where mood of music is predicted with the music features and the training data to construct the prediction model is given by a few volunteers. In another work [29], a method for detecting mood of a text message using fuzzy rules is proposed. However, in this paper, mood tags of multimedia contents given by participants from all over the world are used to constructing a mood fuzzy vector of the multimedia content.

III. A Fuzzy-Based Multimedia Content Retrieval Method Considering Synonyms

Fig. 2 outlines the proposed multimedia content-retrieval method considering synonyms based on fuzzy vectors. The method comprises the following four phases: (1) the construction of multimedia content

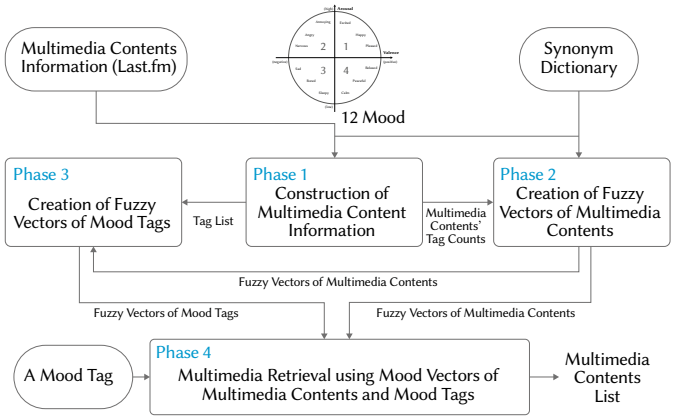


Fig. 2. Multimedia content retrieval structure.

information; (2) the creation of fuzzy vectors of multimedia content; (3) the creation of fuzzy vectors of mood tags; and (4) multimedia content retrieval based on these two fuzzy vectors.

A. Constructing Multimedia Content Information

Fig. 3 presents the process used to construct information for multimedia content divided into two stages: (1) collecting lists of multimedia content on the 12 moods; and (2) collecting tags of multimedia content and tag counts. Although the proposed method can be applied to all types of content, this study considers only music from Last.fm. Since the data collection method is the same as the method of Moon et al. [10], please refer to their work [10] for details.

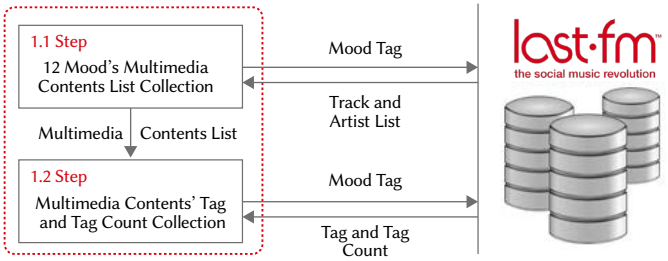


Fig. 3. Multimedia content information construction process.

B. Generating Fuzzy Vectors for Multimedia Content

The process employed to construct fuzzy vectors for multimedia content involves three stages: (1) collecting and analyzing synonyms; (2) calculating mood fuzzy vectors for the multimedia content; and (3) calculating fuzzy vectors of the multimedia content. To classify multimedia content tags into the 12 moods in Fig. 1 to construct a synonym mapping table, the outcome of a previous synonym analysis [1], [10] is used (see Table I).

For generating multimedia content fuzzy vectors, the count values of the 12 moods, called by mood vector, for multimedia content can be calculated using (1) by using the synonym mapping table and relevant tag information as follows:

$$V_k = (M^{1,k}, M^{2,k}, ..., M^{11,k}, M^{12,k})$$
$$M^{m,k} = \sum_{i=1}^{n} t_i^{m,k}$$
$$t_i^{m,k} = \begin{cases} c_i^k, & \text{if } Tag_i^k \in S^m \\ 0, & otherwise \end{cases} \quad (1)$$

where k is the index of the multimedia item, m the mood index ($1 \le m \le 12$), n the number of tags attached to multimedia item k, i the tag index ($1 \le i \le n$), Tag_i^k the i'th tag of multimedia item k, S^m the synonym list (in Table I) corresponding to mood m, and c_i^k the tag count of tag Tag_i^k; $t_i^{m,k}$ is the tag count of tag Tag_i^k for mood m, $M^{m,k}$ the count value of mood m for multimedia item k, and V_k the mood vector of multimedia item k.

For example, assume the tags and counts for the m' th multimedia item in Table II.

TABLE II. Tag and Count

Tag	Count	Tag	Count
annoy	30	**bothersome**	10
angered	15	**Tense**	10
anxious	7	**Excite**	15

TABLE I. Synonym Mapping Table

12 moods	Part	Synonym mood list
annoying	**verb**	annoy, rag, get to, bother, get at, irritate, rile, nark, nettle, gravel, vex, chafe, devil, displease
	adj	bothersome, galling, irritating, nettlesome, pesky, pestering, pestiferous, plaguy, plaguey, teasing, vexatious, vexing, disagreeable
angry	**adj**	aggravated, provoked, angered, enraged, furious, infuriated, maddened, black, choleric, irascible, hot under the collar(predicate), huffy, sore, indignant, incensed, outraged, umbrageous, irate, ireful, livid, smoldering, smouldering, wrathful, wroth, wrothful, raging, tempestuous, stormy, unhealthy, mad, wild
nervous	**adj**	tense, anxious, queasy, uneasy, unquiet, troubled, neural, skittish, flighty, spooky, excitable, aflutter, excited
sad	**adj**	bittersweet, doleful, mournful, heavyhearted, melancholy, melancholic, pensive, wistful, tragic, tragical, tragicomic, tragicomical sad, sorrowful, deplorable, distressing, lamentable, pitiful, sorry, bad
bored	**verb**	bore, tire, drill, cut
	adj	world-weary, tired, blase, uninterested
sleepy	**adj**	sleepy-eyed, sleepyheaded, asleep(predicate)
calm	**verb**	calm down, tranquilize, tranquillise, quieten, lull, comfort, soothe, console, solace, steady, becalm, stabilize, stabilise, cool off, chill out, simmer down, settle down, cool it, sedate, tranquillize, affect, change state, turn
	adj	unagitated, serene, tranquil, composed, placid, quiet, still, smooth, unruffled, settled, windless
peaceful	**adj**	peaceable, irenic, nonbelligerent, pacific, pacifist(prenominal), pacifistic, dovish, peace-loving, halcyon
relaxed	**verb**	relax, loosen, loosen up, unbend, unwind, decompress, slow down, loose, weaken, unstrain, unlax, make relaxed, act, behave, do, slack, slacken, slack up, decrease, lessen, minify, change, alter, modify, affect, change state, turn
	adj	degage, laid-back, mellow, unstrained
pleased	**verb**	please, delight, satisfy, gratify, wish, care, like
	adj	amused, diverted, entertained, bucked up(predicate), encouraged, chuffed, delighted, gratified, proud of(predicate), proud
happy	**adj**	blessed, blissful, bright, golden, prosperous, laughing(prenominal), riant, felicitous, fortunate, glad, willing, well-chosen, halcyon
excited	**verb**	excite, arouse, elicit, enkindle, kindle, evoke, fire, raise, provoke, stimulate, impact, bear upon, bear on, touch on, touch, stir, sensitize, sensitise, agitate, rouse, turn on, charge, commove, charge up, disturb, upset, trouble, sex, wind up, shake, shake up, energize, energise, change, alter, modify, affect
	adj	aroused, emotional, worked up, agitated, agog, crazy, fevered, intoxicated, drunk, overexcited, stimulated, stirred, stirred up, teased, titillated, thrilled, thrillful, delirious, frantic, unrestrained, activated, reactive, aflutter, nervous, mad, wild

Applying the synonyms in Table I to the above data collects "annoy" and "bothersome" (Table II) as synonyms of the "annoying" Thayer mood, giving a count of 40 for this mood. "Angered" is the only synonym of "angry", giving that mood a count value of 15. Likewise, "tense" and "anxious" are synonyms of "nervous", and thus the count of "nervous" is 17. "Excite" is the only term contributing to the count for "excited", which thus has a value of 15. Given that "excited", "annoying", "angry", and "nervous" are, respectively, the 3rd, 4th, 5th, and 6th moods in the Thayer model, the mood vector for multimedia item k is as follows:

$$V_k = (0, 0, 15, 40, 15, 17, 0, 0, 0, 0, 0, 0)$$

Finally, the mood fuzzy vector of a multimedia content is obtained, as indicated in (2). Each dimension value of the vector indicates membership of the mood that is pertinent to the dimension. For instance, the CF_k value of V_k explained above is (0.00, 0.00, 0.17, 0.46, 0.17, 0.20, 0.00, ..., 0.00).

$$CF_k = \left(\frac{V_k^1}{SV_k}, \frac{V_k^2}{SV_k}, \ldots, \frac{V_k^{12}}{SV_k}\right), SV_k = \sum_{i=1}^{12} V_k^i \tag{2}$$

where k is the index of a multimedia content, V_k^i the count value of i' th mood of multimedia content k, SV_k the total mood count value of multimedia content k, and CF_k the mood fuzzy vector of multimedia content k.

In the work by Moon et al. [10], mood count values of 12 moods are converted to an AV value that is used as an internal tag of multimedia content. During the conversion process, the mood information is partially lost, resulting in degrading retrieval performance. For improving retrieval performance, in this paper, the possibility of each mood is calculated by equation (2). Also, considering this possibility, the similarity between the query and the multimedia contents is calculated.

C. Generating Mood Fuzzy Vector for Tags

The mood fuzzy vector of a tag is requisite to search multimedia contents by using mood fuzzy vectors of multimedia contents. The mood fuzzy vector of a tag is calculated by the process shown in Fig. 4. The mood fuzzy table of tags is generated as the outcome of this process. The mood fuzzy vector of a tag is calculated as shown in (3) as the mean of the mood fuzzy vectors of multimedia contents, which includes the mood tag:

$$TF_t = \frac{\sum_{i=1}^{l(t)} CF_i}{l(t)} \tag{3}$$

where TF_t is the fuzzy vector of tag t; $l(t)$ is the number of multimedia contents that includes tag t; and CF_i is the fuzzy vector of content i.

For example, in Fig. 4, the mood fuzzy vector of "Mood Tag (1)" can be calculated by the formula average(CF_1, CF_2, ..., CF_i, $CF_{l(1)}$), where n is the number of multimedia contents; and CF_i is the fuzzy vector of i'th multimedia contents that include "Mood Tag (1)".

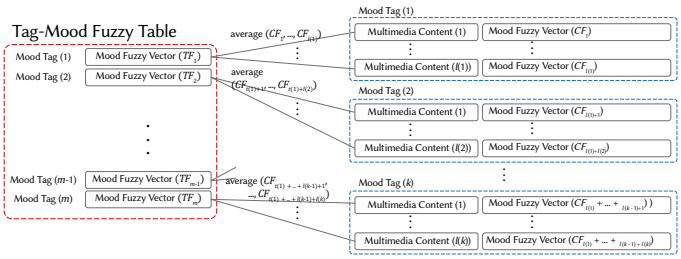


Fig. 4. Tag-mood fuzzy table creation process.

For example, assume the mood fuzzy vector of multimedia content in Table III. From the data, the mood fuzzy vector of the 'happy' tag is generated as follows: the 'happy' tag exists in the multimedia content 1 and 4, and thus the mood fuzzy vector is calculated as {(0.0, ..., 0.22, 0.45, 0.25, 0.07, 0.0) + (0.0, ..., 0.24, 0.63, 0.14)}/2 = (0.00, ..., 0.11, 0.22, 0.24, 0.35, 0.07). The 'sad' tag exists at 2, 5, 8, 9, and 10, so the mood fuzzy vector of the 'sad' tag is calculated as {(0.0, 0.0, 0.0, 1.0, ..., 0.0) + (0.0, 0.32, 0.32, 0.36, ..., 0.0) + (0.0, 0.42, 0.44, 0.14, ..., 0.0) + (0.0, 0.0, 0.17, 0.46, 0.17, 0.2, ..., 0.0) + (0.0, 0.0, 0.27, 0.41, 0.33, ..., 0.0)}/5 = (0.0, 0.15, 0.24, 0.47, 0.10, 0.04, ..., 0.0).

D. Fuzzy-Based Multimedia Content Retrieval

Fig. 5 shows the two stages of multimedia content retrieval based on mood fuzzy vectors of mood tags. The first stage involves searching the mood fuzzy vector of a tag, and the second comprises similarity calculation.

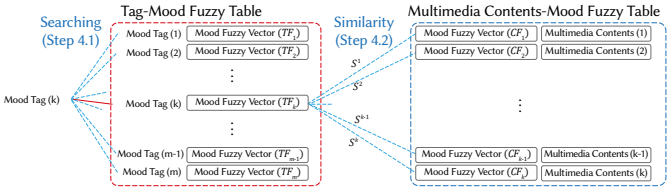


Fig. 5. Multimedia content retrieval method using multimedia content-mood fuzzy table and tag-mood fuzzy table.

After obtaining the mood fuzzy vector of the tag in the user query from the table of mood fuzzy vectors for tags, the similarity between it and the mood fuzzy vector of each multimedia item is calculated, as shown on the right-hand side of Fig. 5. Multimedia content with

TABLE III. Example of the Mood Fuzzy Vector of Multimedia Content

Multimedia Content No	Tag 1	Tag 2	Tag 3	Tag 4	...	Tag n	annoying	angry	nervous	sad	bored	sleepy	calm	peaceful	relaxed	pleased	happy	excited
1	peaceful	relaxed	pleased	tag	...	happy	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.22	0.45	0.25	0.07	0.00
2	sad				...		0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	rag	angry	nervous		...		0.38	0.31	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	amused	happy	excited		...		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.24	0.63	0.14
5	angry	nervous	sad		...		0.00	0.32	0.32	0.36	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	peaceful	relaxed	pleased		...		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.33	0.33	0.00	0.00
7	rag	angry	nervous		...		0.21	0.57	0.21	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	angry	nervous	sad		...		0.00	0.42	0.44	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	sad	sleep	tense	tag	...	asleep	0.00	0.00	0.17	0.46	0.17	0.20	0.00	0.00	0.00	0.00	0.00	0.00
10	sad	tire	nervous		...		0.00	0.00	0.27	0.41	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00

high similarity is preferentially retrieved. Similarities are as shown in Table IV.

TABLE IV. SIMILARITIES AND THEIR EQUATIONS

Similarity	Equation
Similarity of [11]	$s_k(TF_k, CF_l) = \frac{1}{m}\sum^{m}\left(\frac{min(TF_k^i, CF_l^i)}{max(TF_k^i, CF_l^i)}\right)$ (4) where the denominator equalizes to zero, then $s_k(TF_k, CF_l) = 1$.
Similarity of [12]	$s_k(TF_k, CF_l) = \sum_{i=1}^{m}\frac{2 \times min(TF_k^i, CF_l^i)}{TF_k^i + CF_l^i}$ (5) where if $TF_k^i + CF_l^i = 0$, then $s_k(TF_k, CF_l) = 1$.
Similarity of [13]	$s_k(TF_k, CF_l) = \frac{\sum_{i=1}^{m} min(TF_k^i, CF_l^i)}{\sum_{i=1}^{m} max(TF_k^i, CF_l^i)}$ (6)
Similarity of [14]	$s_k(TF_k, CF_l) = \frac{1}{\sqrt{m}}\sqrt{\sum_{i=1}^{m}(TF_k^i - CF_k^i)^2}$ (7)

In the equation, $s_k(TF_k, CF_l)$ means similarity between mood fuzzy vector (TF_k) of tag k and mood fuzzy vector (CF_l) of l'th multimedia content; and TF_k^i and CF_l^i mean i'th value of each mood fuzzy vector.

IV. EXPERIMENTS

To analyze the performance of the proposed method, approximately 50,000 multimedia items and their tags are collected from the API of the website Last.fm. In detail, the dataset contains the following numbers of each tag: 3,200 "angry"; 1,900 "annoying"; 40 "bored"; 10,000 "calm"; 70 "excited"; 10,300 "happy"; 170 "nervous"; 4,800 "peaceful"; 10 "pleased"; 1,900 "relaxed"; 10,200 "sad"; and 7,500 "sleepy". The dataset is used to construct the mood fuzzy tables of the multimedia content and of the tags. Then, the retrieval performance and distribution of the mood fuzzy vectors of the content and tags are analyzed using these tables. The recall and precision are calculated to evaluate the retrieval performance. After obtaining the interpolated precision at 10 recall levels [31], the average precision is calculated for the 12 Thayer mood words. The mood count vectors of multimedia contents and tags and the pseudocodes for calculating the mood fuzzy vectors from them are available at [32].

The recall and precision are calculated by (8), the interpolated precision is calculated by equation (9) and the average precision is calculated by (9).

$$P = \frac{|B|}{|B| + |C|},\ r = \frac{|B|}{|A| + |B|}, \tag{8}$$

where, r is tag recall, P precision, A false negative, C false positive and B true positive.

$$P(r_j) = max_{r_i \le r \le r_{i+1}} P(r) \tag{9}$$

where, $P(r)$ is a precision of recall level r and $r_i \le r \le r_{i+1}$ range of recall level. In this paper, r value range is 0.1.

$$\overline{P(r)} = \sum_{i=1}^{N_q}\frac{P_i(r)}{N_q} \tag{10}$$

where, N_q is number of queries, $P_i(r)$ precision of i'th query in recall r.

A. Analysis of the Mood Fuzzy Vectors of Multimedia Contents and Tags

Prior to analysis of retrieval performance, the fuzzy table of mood tags and the fuzzy table of multimedia contents are analyzed based on 12 mood tags of the Thayer two-dimensional model. According to the analysis, the membership value of the relevant mood of each mood tag (e.g., "pleased" mood for "pleased" tag, "happy" mood for "happy" tag, etc.) is high, as shown in Fig. 6. Among them, the tags "pleased", "excited" and "nervous" have a large membership value of relevant tag mood, and the membership values of the other moods are also large compared to other mood tags. This may be ascribed to the fact that the mood fuzzy vectors for the three tags are calculated by use of fewer than 200 multimedia content information, while more than approximately 1,000 for the residual nine tags.

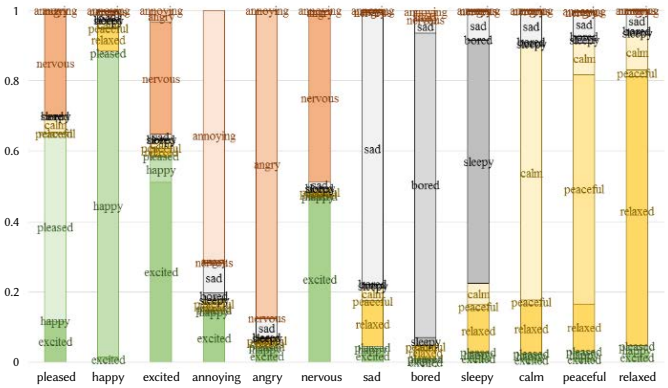
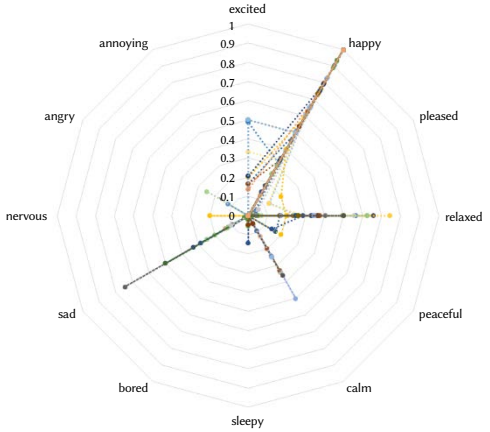
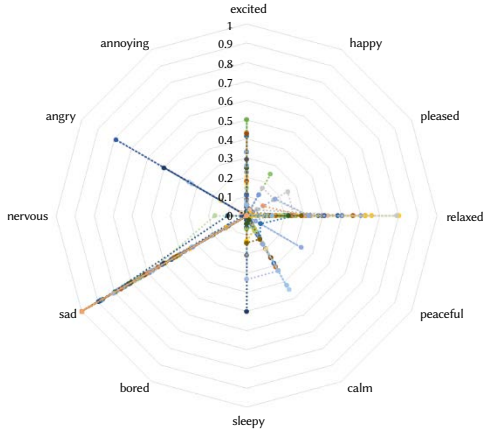


Fig. 6. Multimedia content retrieval method using multimedia content-mood fuzzy table and tag-mood fuzzy table.



(a) Multimedia content that includes the happy tag



(b) Multimedia content that includes the sad tag

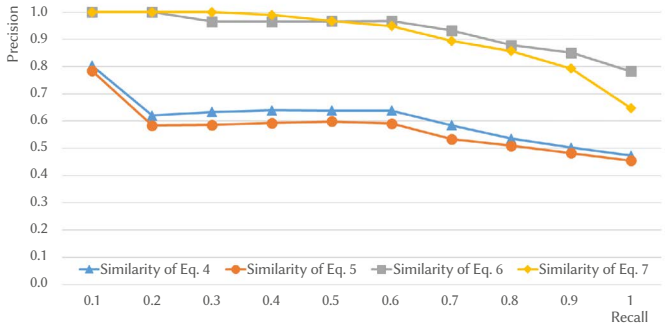
Fig. 7. Analysis of the mood fuzzy table of multimedia content.

TABLE V. ANOVA Test (MFV: Mood Fuzzy Vector)

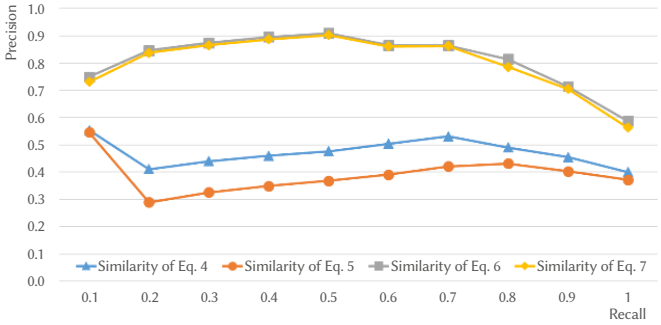
		Sum of Squares	DF	Mean Square	F	P-Value		Sum of Squares	DF	Mean Square	F	P-Value
Mood		700.730	11	63.703	11028.258	.000		3480.844	11	316.440	10382.104	.000
Error	**MFV1**	265.121	45898	.006			**MFV7**	1398.944	45898	.030		
Total		965.851	45909					4879.788	45909			
Mood		2037.450	11	185.223	36203.221	.000		1624.819	11	147.711	10769.730	.000
Error	**MFV2**	234.823	45898	.005			**MFV8**	629.508	45898	.014		
Total		2272.273	45909					2254.326	45909			
Mood		43.396	11	3.945	6167.032	.000		807.645	11	73.422	1984.547	.000
Error	**MFV3**	29.362	45898	.001			**MFV9**	1698.088	45898	.037		
Total		72.758	45909					2505.733	45909			
Mood		3800.000	11	345.455	10222.222	.000		3.038	11	.276	161.449	.000
Error	MFV4	1551.098	45898	.034			**MFV10**	78.505	45898	.002		
Total		5351.098	45909					81.543	45909			
Mood		27.451	11	2.496	4672.943	.000		5439.302	11	494.482	32791.637	.000
Error	MFV5	24.512	45898	.001			**MFV11**	692.120	45898	.015		
Total		51.963	45909					6131.422	45909			
Mood		2686.013	11	244.183	12243.833	.000		68.857	11	6.260	609.643	.000
Error	MFV6	915.360	45898	.020			**MFV12**	471.271	45898	.010		
Total		3601.373	45909					540.128	45909			

TABLE VI. Levene Test (MFV: Mood Fuzzy Vector)

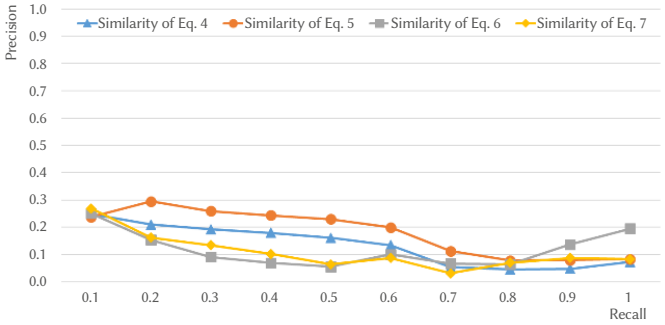
	MFV1	MFV2	MFV3	MFV4	MFV5	MFV6	MFV7	MFV8	MFV9	MFV10	MFV11	MFV12
Levene	10816.91	3610.548	454.9322	1061.525	746.1625	11109.76	4558.702	10067.41	908.7034	54.3630076	3547.89	778.6744
DF1	11	11	11	11	11	11	11	11	11	11	11	11
DF2	45898	45898	45898	45898	45898	45898	45898	45898	45898	45898	45898	45898
P-Value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000



(a) with synonyms



(b) without synonyms



(c) calculated by (a)-(b)

Fig. 8. Retrieval performance comparison by 12 mood tags.

The fuzzy table of multimedia content is analyzed by randomly selecting 200 multimedia contents, including "happy" or "sad", which have the highest number of multimedia contents. The analysis shows that "happy" has a large membership value as shown in Fig. 7 (a), and "sad" also has a large membership value as shown in Fig. 7 (b). For reference, the remaining 10 moods also exhibit a propensity that is similar to that of "happy" and "sad".

This paper performs an ANOVA test and a test for equality of variances (the Levene test) to determine whether they have an independent distribution for mood fuzzy vectors of the fuzzy table of multimedia content. The 12 moods (happy, sad, annoying, pleased, excited, nervous, bored, sleepy, calm, peaceful, relaxed, and angry) of multimedia content are selected as independent variables, and the mood fuzzy vector of each multimedia content is selected as dependent variables. The outcome of the experiment is shown in Tables V and VI. Since null hypothesis H_0 can be rejected because all p-values are 0.000, alternative hypothesis H_1 can be adopted. Specifically, it can be concluded that the difference in distribution of mood fuzzy vectors and the difference in dispersion arise when the mood fuzzy vectors of multimedia content are classified into 12 mood groups.

B. Analysis of Fuzzy-Based Retrieval Performance Using Mood Tags

To analyze the retrieval performance of the proposed method, we perform three kinds of comparisons. The first is the comparison of retrieval performance for four similarities in Table IV, and the second is the comparison of the retrieval performance of the proposed method with the keyword-based method of Last.fm. The last is the comparison of the retrieval performance with the method based on the AV value [10]. The first experiment shows that when analogue is considered, methods using formulas (6) and (7), as shown in Fig. 8 (a), achieve performance of precision 1.0 at recall level 0.1 and 0.2; methods using formulas (4) and (5) achieve performance below precision 0.8 at recall level 0.1 and performance below precision 0.65 at recall level 0.2. Thus, the method that uses formulas (6) and (7) shows better performance than the method that uses formulas (4) and (5) overall. When analogue is not considered, however, the method that uses formulas (6) and (7) also achieves better performance than the method that uses formulas (6) and (7) (refer to (b) and (c)).

When comparing the keyword-based retrieval performance of Last.fm with the retrieval performance of the proposed method, the same performance is achieved with precision 1.0 at recall level 0.1 and 0.2, as shown in Fig. 9. However, the proposed method (the method that uses formulas (6) and (7)) shows better performance than the keyword-based retrieval method from recall level 0.3. Although the keyword-based retrieval performance approaches precision 0.0 when it approaches recall level 1.0, the proposed method remains above precision 0.6 even when it approaches recall level 1.0. Consequently, it can be concluded that the proposed method (the method that uses formulas (6) and (7)) shows better performance than the keyword-based retrieval method overall (refer to Fig. 9).

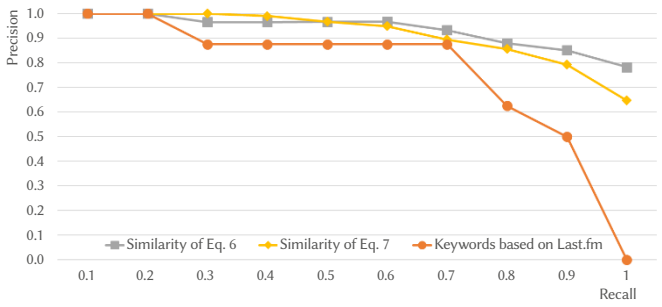


Fig. 9. Retrieval performance comparison between the proposed method and the Last.fm method.

Finally, comparing the retrieval performance of the AV-based method in [10] with that of the proposed method, the proposed method (the method that uses formulas (6) and (7)) achieves excellent performance, as shown in Fig. 10. The gap in retrieval performance is larger than precision of 0.1 at all recall levels. Importantly, the problem of the existing AV-based method at recall level 0.1 and 0.2, i.e., the reduction of retrieval performance, can be solved. In Figure 10, case #1 means retrieval performance in which tags having more than 1,000 multimedia contents are used, and case #2 means retrieval performance in which all 12 mood tags are used.

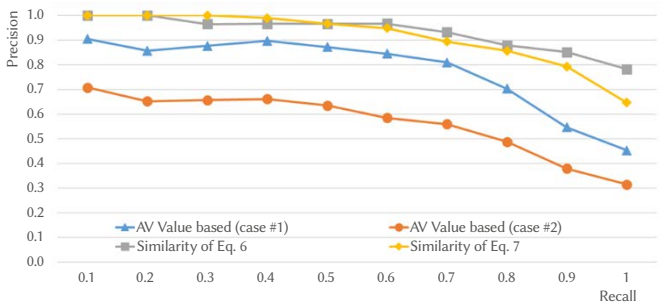


Fig. 10. Retrieval performance comparison between the proposed method and the AV-based method [10].

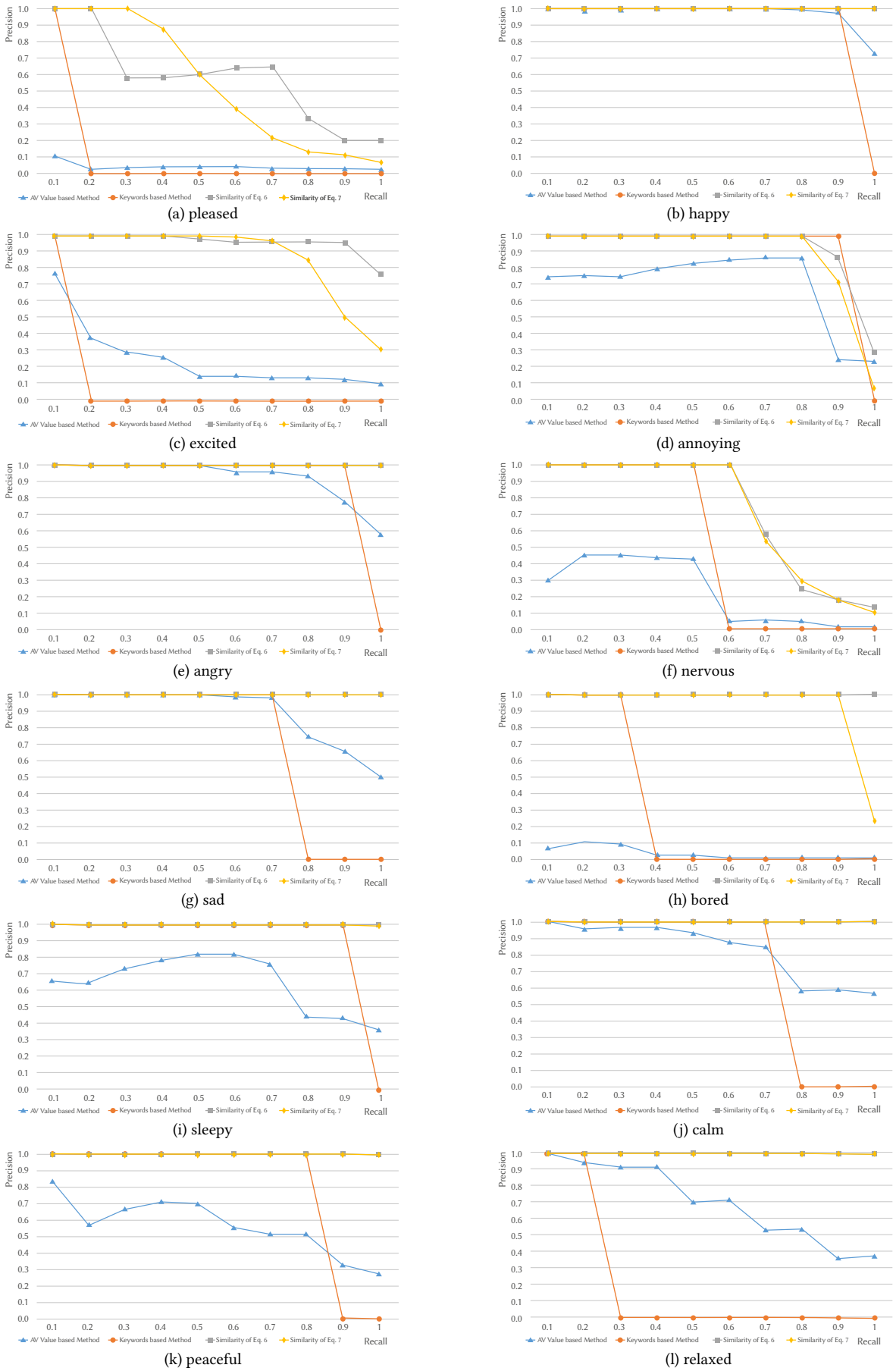
To analyze the retrieval performance in detail, the AV-based retrieval method [10], the keyword-based method of Last.fm, and the proposed method (the method that uses formulas (6) and (7)) are compared and analyzed on the basis of the 12 moods. According to the comparative analysis, the retrieval method of the proposed method shows excellent performance overall, as shown in Fig. 11. Moreover, the proposed method achieves excellent performance even with tags (pleased, excited, bored, and nervous) below 1,000 multimedia contents, for which the AV-based method [10] reveals performance degradation. Overall, the proposed method achieves superior performance compared to the extant methods.

V. Conclusion

The propensity of information purchasers on the Web is changing markedly from cost-effective to cost-satisfaction consumption. One of the methods to maximize cost-satisfaction constitutes utilizing multimedia contents' mood. Some SNSs provide such a service through the use of folksonomy. However, folksonomy presents some problems, such as synonym and new coinage. To solve the synonym problem, a study [10] attempted to represent multimedia content with AV value (arousal and valence) of the Thayer model, but retrieval performance was inferior to the keyword-based method at recall level 0.1.

To overcome the challenge in [10], this paper proposed a method that uses fuzzy moods of multimedia content based on the Thayer mood model. This proposed method solved the synonym problem, as well as achieved retrieval performance that was the same as the keyword-based method at recall level 0.1. The keyword-based retrieval method does not take into account synonym and so suffer from performance degradation from recall level 0.3. In [10], some mood information is lost and thus performance degradation at recall level 0.1 and 0.2 is caused. However, the proposed method does not.

The method proposed in this paper was analyzed in two ways. The first is the analysis of the mood fuzzy table of tags and the mood fuzzy table of multimedia content that constituted intermediate outputs of the proposed method. The analysis demonstrated that the distribution of the mood fuzzy vectors of multimedia content is similar to the distribution of the mood fuzzy vector of tags. In particular, we conducted an ANOVA test and a test for equality of variances (the



(a) pleased (b) happy

(c) excited (d) annoying

(e) angry (f) nervous

(g) sad (h) bored

(i) sleepy (j) calm

(k) peaceful (l) relaxed

Fig. 11. Retrieval performance comparison per mood tag.

Levene test) using the mood fuzzy table of multimedia content. Subsequently, it was found that a difference of mood fuzzy vector distribution and dispersion occurred.

Second, the retrieval performances of the three methods – the keyword-based method of Last.fm, the AV-based retrieval method [10], and the proposed method - were analyzed. The analysis revealed that the problem of the AV-based retrieval method (the retrieval performance of multimedia content below 1,000 at recall level 0.1 and 0.2 is inferior to the retrieval performance of the keyword-based retrieval method) could be solved, and retrieved objects were increased by more than 50% when synonyms are considered. Overall, the method proposed in this paper is far superior to that of the keyword-based method of Last.fm and the AV-based retrieval method.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2020R1F1A104833611, 2021R1I1A1A01042270).

References

[1] C. B. Moon, J. Y. Yi, D.-S. Kim, B. M. Kim, "Analysis of Overlapping Mood Tags Based on Synonyms," *Korea Computer Congress 2018 (KCC 2018), KIISE (2018)*, June 20-22; ICC JEJU, Korea, pp.667-669, 2018, doi: 10.9723/JKSIIS.2019.24.1.013.

[2] Russel, J. A., "A Circumplex Model of Affect," *Journal of Personality and Social Psychology*, Vol.39, No.6, pp.1161-1178, 1980, doi: 10.1037/h0077714.

[3] Hevner, K., "Experimental studies of the elements of expression in music," *The American Journal of Psychology*, Vol.48, No.2, pp.246-268, 1936, doi: 10.2307/1415746.

[4] Thayer, R. E., *The Biopsychology of Mood and Arousal*, Oxford University Press, 1990.

[5] C. B. Moon, H. S. Kim, B. M. Kim, "Music Retrieval Method using Mood Tag and Music AV Tag based on Folksonomy," *Journal of KIISE*, Vol.40, No.9, pp.526-543, 2013, doi: 10.9723/jksiis.2019.24.1.013.

[6] C. B. Moon, H.S. Kim, H. A. Lee, B. M. Kim, "Analysis of Relationships Between Mood and Color for Different Musical Preferences," *Color Research & Application*, Vol.39, No.4, pp.413-423, 2014, doi: 10.1002/col.21806.

[7] C. B. Moon, H.S. Kim, D. W. Lee, B. M. Kim, "Mood Lighting System Reflecting Music Mood," *COLOR research and application*, Vol.40, No.2, pp.201-212, 2015, doi: 10.1002/col.21864.

[8] Viskovich, Shelley Narelle, *Evaluation of a web-based Acceptance and Commitment Therapy (ACT) program to increase self-care and mental health skills in university students*, PhD Thesis, School of Psychology, The University of Queensland, 2019.

[9] Klinenberg, E., *Going solo: the extraordinary rise and surprising appeal of living alone*, London: Penguin Books, 2012.

[10] C. B. Moon, J. Y. Lee, D.-S. Kim, B. M. Kim, "Multimedia Content Recommendation in Social Networks using Mood Tags and Synonyms," *Multimedia Systems*, 26, pp.139–156, 2020, doi: 10.1007/s00530-019-00632-w.

[11] W.-J. Wang, "New similarity measures on fuzzy sets and on elements," *Fuzzy sets and systems*, vol. 85, pp. 305–309, 1997, doi: 10.1016/0165-0114(95)00365-7.

[12] D.-G. Wang, Y.-P. Meng, and H.-X. Li, "A fuzzy similarity inference method for fuzzy reasoning," *Computers and mathematics with applications*, vol. 56, pp. 2445–2454, 2008, doi: 10.1016/j.camwa.2008.03.054.

[13] C. P. Pappis and I. Karacapilidis, "A comparative assessment of measures of similarity of fuzzy values," *Fuzzy sets and systems*, vol. 56, pp. 171–174, 1993, doi: 10.1016/0165-0114(93)90141-4.

[14] L. Baccour, A. M. Alimi, R. I. John, "Some notes on fuzzy similarity measures and application to classification of shapes recognition of arabic sentences and mosaic," *IAENG International Journal of Computer Science*, vol. 41, no. 2, pp. 81-90, 2014.

[15] Huang, W., Wu, Q., Dey, N., Ashour, A., Fong, S. J., & González-Crespo, R., "Adjectives Grouping in a Dimensionality Affective Clustering Model for Fuzzy Perceptual Evaluation," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 10, pp. 28-37, 2020, doi: 10.9781/ijimai.2020.05.002

[16] Ness, S. R., Theocharis, A., Tzanetakis, G. and Martins, L. G., "Improving automatic music tag annotation using stacked generalization of probabilistic svm outputs," *Proc. of the 17th ACM international conference on Multimedia*, pp.705-708, 2009, doi: 10.1145/1631272.1631393.

[17] Laurier, C., Sordo, M., Serra, J. and Herrera, P., "Music mood representations from social tags," *Proc. of the 10th International Society for Music Information Conference*, Kobe, Japan, pp.381-386, 2009.

[18] Kim, J., Lee, S., Kim, S. and Yoo, W. Y., "Music mood classification model based on arousal-valence values," *Proc. of 13th International Conference on Advanced Communication Technology (ICACT)*, 2011, pp.292-295, 2011.

[19] A. T. Ji, et al., "Collaborative Tagging in Recommender Systems," *AI 2007: Advances in Artificial Intelligence*, pp. 377–386, 2007, doi: 10.1007/978-3-540-76928-6_39.

[20] K. H. L. Tso-Sutter, et al., "Tag-aware Recommender Systems by Fusion of Collaborative Filtering Algorithms," *SAC '08: Proceedings of the 2008 ACM symposium on Applied computing*, pp. 1995–1999, 2008, doi: 10.1145/1363686.1364171.

[21] M. Vojnovic, et al., "Ranking and Suggesting Popular Items," *IEEE Transactions on Knowledge and Data Engineering*, vol.21, pp.1133-1146, 2009, doi: 10.1109/TKDE.2009.34.

[22] S. Yang, S. Kim and Y. M. Ro, "Semantic Home Photo Categorization," *IEEE Trans. Circuits and Systems for Video Technology*, Vol.17, No.3, March 2007, pp.324-335, 2007, doi: 10.1109/TCSVT.2007.890829.

[23] S. Yang, S. K. Kim, K. S. Seo, Y. M. Ro, J. Kim and Y. S. Seo, "Semantic categorization of digital home photo using photographic region templates," *International Journal of Information Processing and Management*, Vol.43, No.2, March 2007, pp.503-514, 2007, doi: 10.1016/j.ipm.2006.07.009.

[24] S. Yang, R. M. Ro, "Photo Indexing Using Person-Based Multi-feature Fusion with Temporal Context," *Proc. International Conference on Mobile Ubiquitous Computing, Systems*, Services and Technologies, Nov. 2007, pp.257-262, 2007, doi: 10.1109/UBICOMM.2007.29.

[25] E. Chang, G. Kingshy, G. Sychay, G. Wu, "CBSA: Content-Based Soft Annotation for Multimodal Image Retrieval using Bayes Point Machines," *IEEE Trans. Circuits and Systems for Video Technology*, Vol.13, No.1, January 2003, pp.26 – 38, 2003, doi: 10.1109/TCSVT.2002.808079.

[26] J. Li and J. Z. Wang, "Real-Time Computerized Annotation of Pictures," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.30, No.6, June 2008, pp.985-1002, 2008, doi: 10.1109/TPAMI.2007.70847.

[27] R. M ö rzinger, R. Sorschag, G. Thallinger and S. Lindstaedt, "Automatic Image Annotation using Visual Content and Folksonomies," *Multimedia Tools and Applications*, Vol.42, No.1, March 2009, pp.97-113, 2009, doi: 10.1007/s11042-008-0247-7.

[28] S.h. Rho, B.-J. Han, and E. Hwang, "A fuzzy inference-based music emotion recognition system," *In Proc. Visual Information Engineering*, pp. 673–677, 2008, doi: 10.1049/cp:20080398.

[29] S. Qamar, P. Ahmad, "Emotion Detection from Text using Fuzzy Logic," *International Journal of Computer Applications*, Vol. 121, No. 3, pp. 29-32, 2015, doi: 10.5120/21522-4501.

[30] Magdin, M., Držík, D., Reichel, J., & Koprda, S., "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, doi: 10.9781/ijimai.2020.11.010

[31] D. Powers, "Evaluation: From precision recall and f-measure to roc informedness markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37-63, 2011, doi: 10.9735/2229-3981.

[32] https://drive.google.com/file/d/18KmHezs55_fMNj95XgceIDYOhhxu3HmZ/view

Chang Bae Moon

Chang Bae Moon received a BSc, an MSc, and a PhD from the Dept. of Software Eng. at Kumoh National Institute of Technology, Korea, in 2007, 2010, and 2013, respectively. He has been with the Kumoh National Institute of Technology since 2014 as a Research Professor in the ICT Convergence Research Center. From 2013 to 2014, he was a Senior Researcher in Young Poong Elec. Co. His current research areas include artificial intelligence, Web intelligence, information filtering, and image processing.

Jong Yeol Lee

Jong Yeol Lee received the BS and MS degree in Dept. of computer Eng. from Kumoh National Institute of Technology, Korea, in 1992 and 1994, respectively and the PhD candidate in software Eng. from Kumoh National Institute of Technology, Korea, in 2018. He has been with Kumoh National Institute of Technology since 2005 as a time lecturer of Computer Software Engineering Department. His current research areas include Artificial Intelligence, Machine Learning and Information Security.

Byeong Man Kim

Byeong Man Kim received the BS degree in Dept. of computer Eng. from Seoul National University (SNU), Korea, in 1987, and the MS and the PhD degree in computer science from Korea Advanced Institute of Science and Technology (KAIST), Korea, in 1989 and 1992, respectively. He has been with Kumoh National Institute of Technology since 1992 as a faculty member of Computer Software Engineering Department. From 1998 – 1999, he was a post-doctoral fellow in UC, Irvine. From 2005 - 2006, he was a visiting scholar at Dept. of Computer Science of Colorado State University, working on design of a collaborative Web agent based on friend network. His current research areas include artificial intelligence, Web intelligence, information filtering and brain computer interface.

A Rule-Based Expert System for Teachers' Certification in the Use of Learning Management Systems

Luisa M. Regueras, María Jesús Verdú*, Juan-Pablo de Castro

Higher Technical School of Telecommunications Engineering (ETSIT), Universidad de Valladolid, Valladolid (Spain)

Received 11 August 2022 | Accepted 7 November 2022 | Published 28 November 2022

Abstract

In recent years and accelerated by the arrival of the COVID-19 pandemic, Learning Management Systems (LMS) are increasingly used as a complement to university teaching. LMS provide an important number of resources and activities that teachers can freely select to complement their teaching, which means courses with different usage patterns difficult to characterize. This study proposes an expert system to automatically classify courses and certify teachers' LMS competence from LMS logs. The proposed system uses clustering to stablish the classification scheme. From the output of this algorithm, it defines the rules used to classify courses. Data registered from a university virtual campus with 3,303 courses and two million interactive events have been used to obtain the classification scheme and rules. The system has been validated against a group of experts. Results show that it performs successfully. Therefore, it can be concluded that the system can automatically and satisfactorily evaluate and certify the teachers' LMS competence evidenced in their courses.

Keywords

Academic Analytics, Automatic Course Classification, Learning Management System (LMS), Rule-Based Expert System.

DOI: 10.9781/ijimai.2022.11.004

I. Introduction

LEARNING Management Systems (LMS) are widely used across universities to support teaching and learning. LMS are a driving force in online courses; however, they are increasingly used as complement to face-to-face classes [1], [2], and much more with the advent of the COVID-19 pandemic and the new educational needs.

LMS provide a wide variety of tools (communication, skills, and productivity) as well as reports of learning progress [3]. Teachers choose the resources and activities that best suit their needs and the way they teach. Thus, in an institutional LMS, there are courses with different usage patterns difficult to characterize [4], [5].

In this context, many universities want to know the usage given by teachers to certificate and evaluate their competence in technology-based learning. This task is done manually and subjectively by experts based on the presence/absence of LMS activities and resources. This is a hard and difficult task; therefore, it would be interesting to automatize this certification process and define an expert system that was able to certificate courses based on the use of LMS by teachers and students.

Moreover, supporting teachers with feedback about their LMS usage could be a motivating factor to make the most of LMS as well as improve their learning designs [6]. This could be also an interesting point, since the lack of motivation to integrate technology is one of the biggest challenges to the implementation of blended teaching [7]. Some studies indicate that the most common use of LMS is as a repository [8]. However, it should be desirable to further utilize LMS capabilities, especially those tools that improve student interaction and increase engagement [9].

This paper presents and validates an expert system that automatically classifies courses and certifies the teachers' competence in the use of LMS from scarce and partial data. For that end, it is previously necessary to stablish the different classes of courses according to the LMS usage [10]. The definition of course types could be done manually by some agent (educational authorities, for example) or automatically by means of some type of clustering analysis. We propose a whole automatic system that firstly applies clustering to categorize courses, like described in our previous work [10], and then estimates the courses typology based on the use of LMS.

In section II, a review of the related work is presented. Section III describes the expert system, along with the used methods and tools. In section IV, the results of this study and the accuracy of the estimations are presented and discussed. Finally, section V contains outcomes and insights about future work.

II. State of Art

Expert systems improve decision-making processes by reasoning through accumulated experience together with an inference or rules engine. Most expert systems are rule-based reasoning, where the knowledge base is represented using rules in the form of IF-THEN. For example, some studies use a forward chaining method to support learning assessment and to assist new-comer students [11], [12]; whereas Hossain et al. [13] develop a belief rule base, an extended form of traditional IF-THEN rules, to predict the student performance under uncertainty. Other authors incorporate fuzzy logic in the expert system to improve students' learning performance [14].

* Corresponding author.
E-mail address: marver@tel.uva.es

In the last years, expert systems for educational purposes are increasingly demanded [15]. Applied on LMS, most expert systems are Intelligent Tutoring Systems [16]. They are recommenders or feedback systems that deal with learning adaptation or personalization [17]–[22]. There are also systems based on students' drops or outcomes prediction and learning problem solving [23], [24].

On the other hand, very few systems are addressed to the teachers' view or improvement [25]–[27]. For example, Hossain et al. [28] describe a data mining tool that supports teachers in making decisions about how to improve their e-learning courses. Villagrá-Arnedo et al. [29] propose a system that provides time-dependent predictions of students' performance so that teachers can select the best moment for intervention. Moreover, expert systems for higher level functions, such as outcomes-based teaching planning [30] or assessment of academic credentials and competencies [31] are also few.

Finally, there are several studies on the characterization of courses according to the level of LMS usage, as we detailed in our earlier work [10]. Previous studies [32]–[34] classified courses from LMS interactions with the aim of supporting instructors in the development of course plans, improving designs or increasing the impact of LMS use. To our knowledge, only the results of Whitmer et al. [34] have been used for actual support systems. Their "course archetypes" have been incorporated into the learning analytics product of Blackboard LMS [35], a proprietary software. However, the implementation and design of that classification system is not provided and presented in the literature. Moreover, their archetypes are defined from courses previously selected, and not from all courses offered. Only courses incorporating the gradebook have been used in the analysis [34], which, as it has been verified in our previous study [10], is quite limited in blended learning environments. Other issue is the great variability of contexts: they work with courses from 60 minutes long (a short workshop) to a whole semester or year. In fact, they use later their course archetypes for a correlation analysis of students' grades and course patterns at a single university and they find unexpected results [36]. Caglayan et al. [6] use also the archetypes provided by Blackboard Analytics to investigate the degree of agreement between instructors' opinion on their course type and the classification done by Blackboard Analytics. The experiment is also at university level and face-to-face classrooms. Their results show a low level of consistency between instructors' view and the analytics findings. However, they conclude that knowing the automatically labelled archetype helps instructors to think about and redesign their courses [6]. In any case, it is another proof of the need of more customized classifications that consider the instructional and cultural context as well as some type of validation.

Other recent studies about characterization of LMS courses can be found in the literature. Machajewski et al. [4] use latent class analysis to characterize courses at a university; whereas Su et al. [27] analyse the behavioural patterns of university teachers while using an LMS. Both find three distinct patterns or clusters but any of them use their findings to give feedback or feed an expert system. Finally, Bennacer et al. [37] are developing a self-assessment tool based on a teacher behavioural model, but this model is not automatically obtained. Instead, they do a mixed analysis that includes a quantitative LMS analysis as well as a qualitative one with some interviews to pedagogical experts. Besides, it is at a very early state.

III. The Expert System

We have designed a rule-based expert system that qualifies the teacher's competence in Moodle and establishes how the teacher makes use of the Learning Management System. To that end, it is necessary to first establish the different types of LMS uses. Subsequently the rules that define the expert system can be designed from expert knowledge or by learning from real data. A data-driven design has the advantages of working automatically and objectively while avoiding experts doing a hard manual work [10].

The process takes place in two phases (see Fig. 1). During the first phase, the clustering system learns from students' and teachers' activity logs and the classification rules and facts base are defined. In the second phase, the expert system infers the certification of teachers' Moodle competence for each course from the obtained rules and facts base.

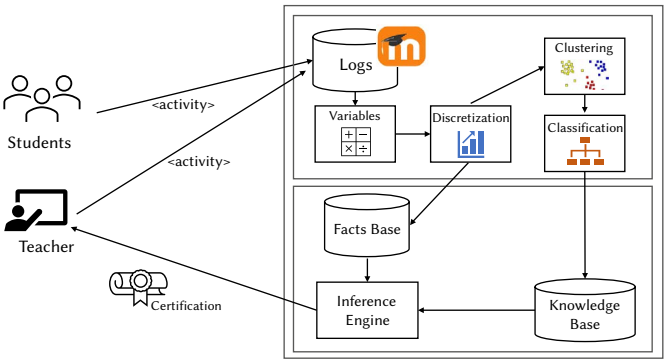


Fig. 1. Architecture of the Expert System.

A. Methods for Clustering Analysis

We used clustering analysis to automatically identify different types of courses in accordance with the LMS usage patterns [10].

From the Moodle logs for all courses given at the University, a process of transformation and selection of variables was conducted to stablish the input variables to the clustering algorithm.

TABLE I. Description of Initial Variables [10]

Variable	Description	Role
Resources	Number of resources	Teacher
ResourceViews	Number of resource views or downloads	Student
Forums	Number of discussion forums	Teacher
ForumNews	Number of teachers' forum posts	Teacher
ForumInteractions	Number of students' forum views and posts	Student
Assigns	Number of assignments	Teacher
AssignSubmissions	Number of assignment submissions	Student
Quizzes	Number of quizzes	Teacher
QuizSubmissions	Number of quiz submissions	Student
OtherActivities	Number of other activities	Teacher
OtherActivitySubmissions	Number of other activity submissions	Student
GradeItems	Number of gradebook items	Teacher
GradeFeedbacks	Number of feedbacks	Teacher
CalendarEvents	Number of manual calendar events	Teacher

A Moodle course can integrate both resources (such as files, links, pages...) and activities (for example, forums, assignments, quizzes, glossaries...). It is also possible to configure tools such as the events calendar and the gradebook for management purposes. According to the possible interactions, 14 variables were selected. Instead of considering items separately, we grouped some of them, especially those with limited use. We only selected the three activities with a more extensive use, and we grouped the rest in another variable as

well as all types of resources, as it is detailed in our previous work [10]. Table I shows the description of variables as well as who carries out the corresponding action (teacher or student), where all activity-related variables are normalized to the number of students enrolled on the course. From these 14 variables, a process of preselection was carried out, by removing the redundant features and the zero and near-zero variance predictors and, therefore, keeping the features of interest.

Once the attributes were selected, the data were discretized to significantly reduce the number of possible values of the variables. We used an unsupervised method (k-means clustering) with three intervals and labels (low, medium, and high) for all variables. From transformed data, we compared different methods of clustering. Finally, we chose LCA (Latent Class Analysis) as clustering method and BIC (Bayesian Information Criterion) as a good indicator of the number of latent classes.

More details about data collection and selection, data pre-processing and transformation and clustering methods are available in our previous work [10].

B. Generation of Rules

Once stablished the different types of courses, we employed decision tree learning to extract information of the courses. This technique visually and explicitly provides a decision representation by deploying a recursive partitioning.

In the field of machine learning, there are different ways to obtain decision trees. We used CART (Classification And Regression Trees), which is a supervised learning technique to obtain classification as well as regression trees. Therefore, we have a target or dependent variable (the course type) and our goal is to obtain a function that allows us to predict, from independent variables, the value of the course type variable. As long as our target variable was discrete, we used the classification variant of CART. We employed the R implementation of CART that is known as Recursive Partitioning and Regression Trees or RPART (the R package 'rpart'). This algorithm finds the independent variable that best separates our data into groups, which correspond to the categories of the target variable. This best separation is expressed with a rule, where each rule corresponds to a node. The main advantage of this method is its interpretability, since it provides a set of rules from which decisions can be made.

We randomly partitioned the data into a training dataset (70%), used to prepare the model, and a test dataset (30%), used to evaluate the model performance, by using data splitting. We measured the goodness of fit of the proposed model by generating a confusion matrix or contingency table, through the 'confusionMatrix' function of caret R package. We repeated the process by changing the training and test dataset to compare the results. Then, we could select the best model depending on our objectives and what accuracy and specificity we wanted to obtain in our predictions.

Finally, once the classification model was chosen, we could use these results to obtain the rules through 'rpart.rules' function and validate their accuracy.

C. Definition of the Expert System

We used CLIPS (C Language Integrated Production System) to define the expert system, since it provides a complete environment for the construction of rule-based expert systems and it is widely used in the definition of expert systems [12], [38]. CLIPS is a forward-chaining rule-based language based on the Rete algorithm for pattern-matching to determine which rule should be fired by the inference engine.

In CLIPS, we defined the template associated to the courses with their features (input to the expert system) and type (output from the expert system). We also implemented the rules that define the system and that had been obtained from the decision tree. From the selected features of each course, the expert system defined in CLIPS obtained the type of course. Next step was to validate the obtained output with the experts' opinion.

IV. Results and Discussion

The hypothesis to be tested in this paper is that the designed expert system performs as a human expert in the task of certificating the teachers' technology competence about the use of LMS.

To analyse and validate the performance of the expert system, it was implemented in CLIPS and tested with real data.

A. Experiment

The study was carried at the Virtual Campus of the University of Valladolid, a Spanish public university, which offers more than 3,000 face-to-face courses. It has more than 2,000 teachers and around 32,000 students enrolled each academic year. This institution has its own virtual campus, based on Moodle LMS, which is being used as a support to face-to-face classes since 2009. All courses have a corresponding course in Moodle, on which both teachers and students are automatically enrolled. However, how to use the platform is decided by each teacher, resulting in different manners and intensities. In this context, the university could be interested in classifying the courses according to LMS usage by using an expert system that would replace manual certification of teachers' on-line competence.

From the anonymized logs, after applying the methods for data pre-processing described in Section III.A, nine variables were selected: 'Resources', 'ResourceViews', 'Forums', 'ForumNews', 'ForumInteractions', 'Assigns', 'AssignSubmissions', 'GradeItems' and 'GradeFeedbacks'. Then, we discretized them using k-means cut-off thresholds. After applying LCA for these variables, six classes were found (see Table II). See our previous work [10] for more details.

TABLE II. Description of Course Types [10]

Type of Course	Description
Type I or Inactive	Low use of Moodle
Type S or Submission	Some content and considerable use of assignments
Type R or Repository	A lot of content and low student interaction
Type C or Communicative	High interaction teacher-students
Type E or Evaluative	Some content and considerable use of evaluative elements
Type B or Balanced	Considerable and balanced use of Moodle tools

Then, decision tree learning was employed to extract information of the courses, as shown in Fig. 2. In this figure, each of the rectangles represents a node of the tree, with its classification rule, the proportion of cases belonging to each category (B C E I R S), and the proportion of the total data that have been grouped there. Each node is coloured according to the category predicted by the model for that group, following the greatest proportion within each region. These proportions give us an idea of the accuracy of the model in making predictions.

We repeated the process by changing the training and test dataset to compare the results and to obtain the best model in accordance with the accuracy, sensitivity (true positive rate) and specificity (true negative rate) shown in the confusion matrix. A confusion matrix is a very useful tool for calibrating the performance of a model and evaluating all possible outcomes of the predictions. Results of Fig. 3 show that the model has a high accuracy (95.8%), and high sensitivity and specificity for the six classes. The worst results were obtained for class B sensitivity, which indicates that the type B is the most difficult to identify correctly by the model.

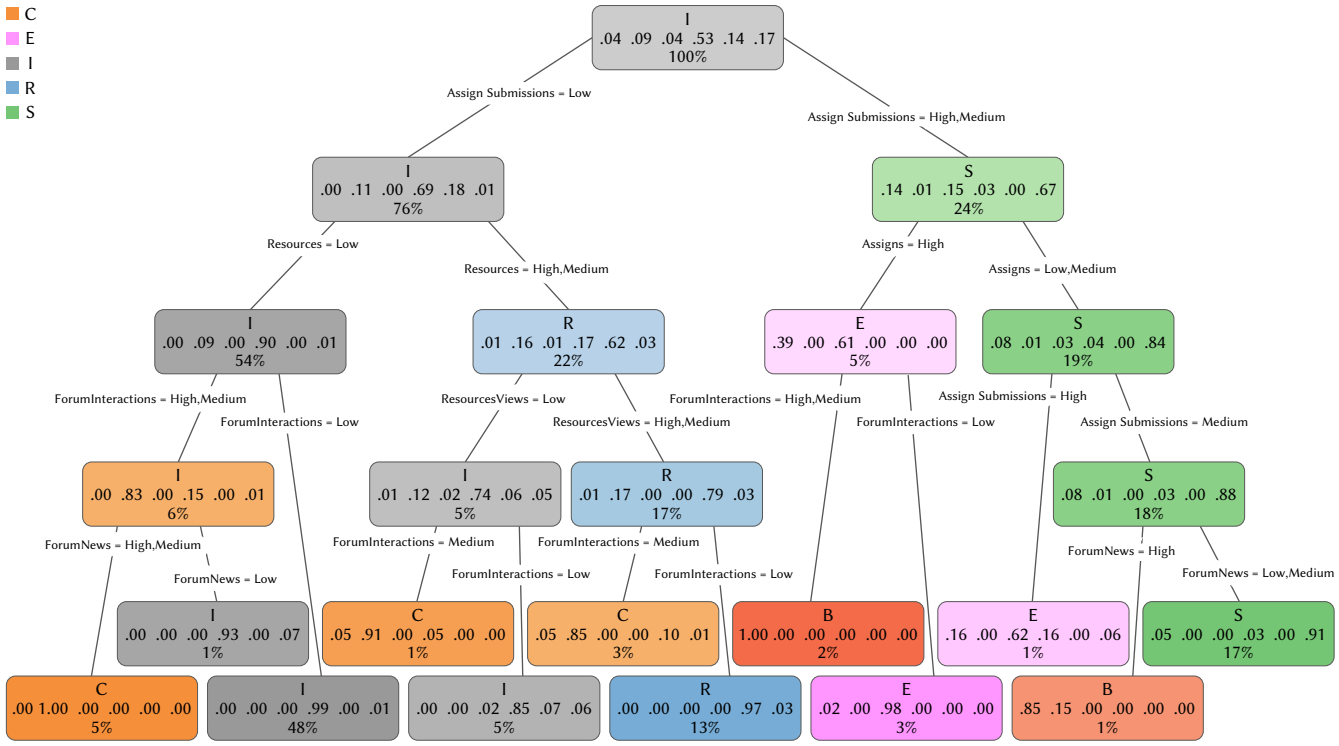


Fig. 2. Decision tree.

```
>>Contusion Matrix and Statistics
          Reference
Prediction    B    C    E    I    R    S
         B   71    3    0    0    0    0
         C    6  260    0    1   10    1
         E    7    0  107    5    0    2
         I    0    0    4 1600   10   21
         R    0    0    1    0  398   12
         S   25    2    0   18    0  482
Overall Statistics
               Accuracy : 0.958
                 95% CI : (0.9502 , 0. 9648)
    No Intormation Rate : 0.5332
    P-Value [Acc > NIR] : < 2.2e-16
                  Kappa : 0.9359
 Mcnemar's Test P-Value : NA

Statistics by class:
                     Class:B   Class:C   Class:E   Class:I   Class:R   Class:S
Sensitivity          0.65138   0.98113   0.95536   0.9852    0.9522    0.9305
Specificity          0.99898   0.99353   0.99523   0.9754    0.9951    0.9822
Pos Pred Value       0.95946   0.93525   0.88430   0.9786    0.9684    0.9146
Neg Pred Value       0.98721   0.99819   0.99829   0.9830    0.9924    0.9857
Prevalence           0.03578   0.08700   0.03677   0.5332    0.1372    0.1701
Detection Rate       0.02331   0.08536   0.03513   0.5253    0.1307    0.1582
Detection Prevalence 0.02429   0.09127   0.03972   0.5368    0.1349    0.1730
Balance Accuracy     0.82518   0.98733   0.97529   0.9803    0.9736    0.9564
```

Fig. 3. Confusion Matrix.

We obtained the rules associated to each model to compare them. We could verify that, while the obtained accuracy and statistics were somewhat different in each repetition, the rules obtained were the same ones, which shows the stability of the rules. From these rules, the expert system was defined: first, the facts template with course features and later, the associated set of rules.

The facts template includes the formal definition of the data. It provides the structure with the names and type associated to all data fields or slots. The course template contains, besides the course identifier, nine slots with the features used to obtain the class and another slot for storing the resultant type (see Fig. 4).

The inference engine of the expert system consists of a set of IF-THEN rules, such as the one shown in the example of Fig. 5. The set of rules is a direct mapping of the decision tree obtained in R (see Fig. 2). Therefore, a total of 12 rules were defined; since, there were types of courses defined with only one rule (types R and S) and other ones with two or three rules (types I, C, E and B).

```
;;*****************
;;* INITIAL STATE *
;;*****************

> (deftemplate course
          "Template far the description of the course"
          (slot IdCourse (type INTEGER))
          (slot Resources (type SYMBOL))
          (slot ResourceViews (type SYMBOL))
          (slot Forums (t ype SYMBOL))
          (slot ForumNews (type SYMBOL))
          (slot Foruminteractions (type SYMBOL))
          (slot Assigns ( t ype SYMBOLJJ
          (slot Assignsubmissions (type SYMBOL))
          (slot Gradeitems (type SYMBOL))
          (slot GradeFeedbacks (type SYMBOL))
          (slot Type (type SYMBOL))
)
```

Fig. 4. Facts template for the expert system.

```
(detrule course_repository
          "Type Repository
          ?f <- (course (IdCourse ?id)
                (Resources ~Low) (ResourceViews ~Low)
                (AssignSubmissions Low) (ForumInteractions Low))
          =>
          (modity ?t (Type R))
          (printout t ?id "is Type R" crlf)
)
```

Fig. 5. Example of rule for the expert system.

B. Validation

The chosen method for validation was "validation against a group of experts" based on the one of Mosqueira-Rey et al. [39], which we had used before successfully [40]. This method provides a measure of agreement between the human experts and verifies if the expert system performs as one of them. In that case, it can be incorporated into the group of experts keeping the agreement level.

TABLE III. EXPERT SYSTEM VS. HUMAN EXPERTS

	Human Expert 1	Human Expert 2	Human Expert 3	Human Expert 4	Human Expert 5
Expert System	I S R C E B	I S R C E B	I S R C E B	I S R C E B	I S R C E B
Inactive – I	20 0 0 0 0 0	20 0 0 0 0 0	20 0 0 0 0 0	9 0 11 0 0 0	20 0 0 0 0 0
Submission – S	0 16 0 0 0 4	0 11 0 0 3 6	0 13 0 0 3 4	0 9 0 0 4 7	0 13 0 0 3 4
Repository – R	0 0 20 0 0 0	0 0 20 0 0 0	0 0 20 0 0 0	0 0 20 0 0 0	0 0 20 0 0 0
Communicative – C	0 0 0 19 0 1	0 0 0 16 0 4	0 0 0 16 0 4	0 0 0 19 0 1	0 0 0 15 0 5
Evaluative – E	0 0 0 0 18 2	0 0 0 0 18 2	0 0 0 0 18 2	0 0 0 0 18 2	0 0 0 0 18 2
Balanced – B	0 0 0 0 0 20	0 1 0 0 0 19	0 1 0 0 0 19	0 4 0 0 0 16	0 0 0 0 0 20

TABLE IV. VALUES OF WEIGHTED KAPPA

	Human Expert 1	Human Expert 2	Human Expert 3	Human Expert 4	Human Expert 5	Expert System
Human Expert 1	–	0.89	0.87	0.78	0.93	0.93
Human Expert 2	0.89	–	0.98	0.79	0.94	0.84
Human Expert 3	0.87	0.98	–	0.77	0.92	0.86
Human Expert 4	0.78	0.79	0.77	-	0.77	0.71
Human Expert 5	0.93	0.94	0.92	0.77	-	0.86
Expert System	0.93	0.84	0.86	0.71	0.86	-

In this experiment, the group of experts consisted of five experts on LMS usage, that is, they were at the same time both teachers with a high level of experience using LMS and researchers in LMS. Since it involved a large workload for one person to evaluate all the courses used to establish the expert system, 20 courses of each typology (120 courses in total) were randomly selected. To classify courses, the experts analysed both the course structure in Moodle and the teachers and students' activity recorded in logs by considering the description of the types of courses obtained above (see Table II). The experts were also asked to talk about the difficulty when classifying the courses and to give their opinion on the quality of the classification scheme.

Table III shows the agreement among the human experts and the expert system for the 120 courses. First column indicates the class obtained by the expert system, 20 courses of each typology, and the other five columns incorporate the classification given by each of the five experts.

Firstly, we can observe that, in general, the human experts agreed with the expert system, especially in the types of courses classified by the expert system as I, R, B and E, and somewhat less in types S and C. Moreover, in most cases, when the human experts differed from the expert system, they classified the course as B type. Therefore, results show also that the type B is the most difficult to identify correctly by the expert system, which is coherent with the sensitivity for B class, as observed in Fig. 3.

The level of agreement between each pair of experts (humans and system) was measured through the weighted kappa [39], which most often deals with data resulting from a judgement. Values of kappa higher than 0.80 indicate an almost perfect agreement whereas values in the range 0.61–0.80 indicate a significant agreement [41]. The results of the measure of kappa (see Table IV) show a good agreement between the expert system and the group of experts. The level of agreement between the expert system and each human expert varies from a significant agreement (kappa = 0.71, with Expert 4) to an almost perfect agreement (kappa = 0.93, with Expert 1), similar to the level of agreement obtained between the human experts.

Fig. 6 shows a heatmap with the level of agreement between the experts and the expert system to analyse visually the results of Table IV. Here, we can see a high degree of agreement provided by the expert system, although it is not perfect. We can also check how there is no total agreement even among the human experts themselves, with expert 4 being the one with the lowest level of agreement with the rest of the experts. This shows how complicated it is to classify courses and how the existence of a system that automates the process is an important advance. Moreover, this is connected to the difficulty and the time spent by experts to classify the courses, since they related that it is not a trivial task.

Finally, it is important to comment the opinion of the experts on the quality of the classification scheme defined by the expert system. They thought that it was very useful and in tune with a subjective analysis of the courses. In addition, they valued positively that it was not a gradual classification but different types of use.

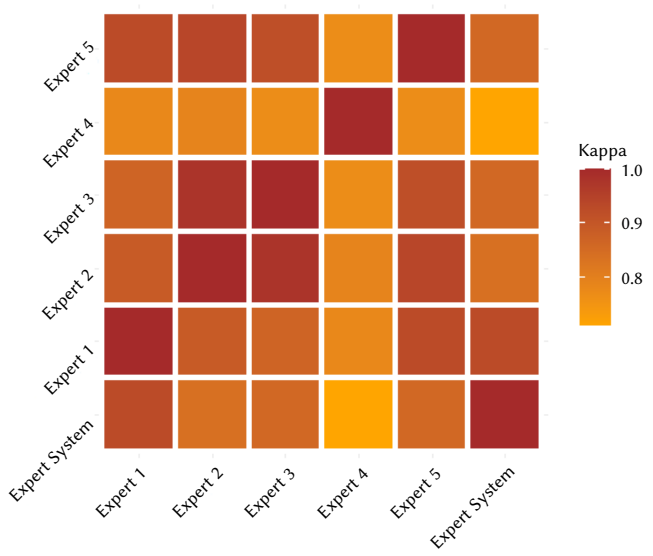


Fig. 6. Heatmap of agreement between experts and the expert system.

V. CONCLUSIONS

We propose an expert system that satisfactorily classifies the courses according to their usage, by both teachers and students. The expert system estimates the typology of the courses according to the real LMS use of students and teachers. The system has been tested with real data and the results have been successfully validated against human experts.

The information provided by the expert system can also be used for reinforcing teachers' continuance commitment to e-learning,

when the perceived self-efficacy is not enough [7], [26]. Besides, it could be adapted to be used by academic administrators in some type of professional career development program, what, in fact, could contribute to Academic Analytics adoption for improving the learning environments, as suggested by some researchers [42], [43].

A generalized comment from the experts was how difficult it was for them to manually classify many of the courses, so they valued very positively that there was a tool capable of automating this process.

This clustering based expert system could support human experts for quick identification of different course types, providing an understanding of how they differ, just like it occurs in other fields [44]. Atypical usages could be also identified for further study.

The experiment may have some limitations that would need to be addressed in further research. For example, many universities offer mainly face-to-face courses and the LMS is only a complement to the face-to-face methodologies. This is an important source of data noise and bias when trying to qualify the online usage of the LMS. Another limitation is the hidden relationship between course categories and time-dependent patterns of events. This experiment only addresses aggregated features over one semester, while it would be valuable to measure the temporal distribution of some features. Some researchers suggest that exploiting time-dependent nature of learning data is both viable and desirable [29]. Moreover, some important subjective aspects such as students' motivation and satisfaction have shown to be sensitive to temporal patterns [2]. These could be used to obtain a better characterization of courses by finding correlations between the structural organization and the emotional effect caused on students.

Future expert system should also incorporate teachers' opinion about the typology of their courses, since they have the best understanding of their learning goals and contexts [6]. Teachers' input could be used to adjust the model.

Finally, a complete intelligent system could incorporate a module to recommend teachers best practices according to the way in which they would like to teach. Teachers would select a course typology. Then, the system would make a proposal about which tools they should use, and some best practices observed in courses of that type with good students' performance and satisfaction.

References

[1] G. Y. Washington, "The Learning Management System Matters in Face-to-Face Higher Education Courses," *Journal of Educational Technology Systems*, vol. 48, no. 2, 2019.

[2] M. Cantabella, R. Martínez-España, B. López, and A. Muñoz, "A Fine-Grained Model to Assess Learner-Content and Methodology Satisfaction in Distance Education," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 4, pp. 87-96, 2020, doi: 10.9781/ijimai.2020.09.002.

[3] N. Nadirah, M. Kasim, and F. Khalid, "Choosing the Right Learning Management System (LMS) for the Higher Education Institution Context: A Systematic Review," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 11, pp. 55–61, 2016.

[4] S. Machajewski, A. Steffen, E. Romero Fuerte, and E. Rivera, "Patterns in Faculty Learning Management System Use," *TechTrends*, vol. 63, no. 5, pp. 543–549, Sep. 2019, doi: 10.1007/s11528-018-0327-0.

[5] Y. Park and I.-H. Jo, "Using log variables in a learning management system to evaluate learning activity using the lens of activity theory," *Assessment & Evaluation in Higher Education*, vol. 42, no. 4, pp. 531–547, May 2017, doi: 10.1080/02602938.2016.1158236.

[6] E. Caglayan, O. O. Demirbas, A. B. Ozkaya, and M. Sahin, "Evidence-Based Learning Design Through Learning Analytics," in *Adoption of Data Analytics in Higher Education Learning and Teaching*, D. Ifenthaler and D. Gibson, Eds. Cham: Springer International Publishing, 2020, pp. 407–424. doi: 10.1007/978-3-030-47392-1_21.

[7] A. Antwi-Boampong, "Towards a faculty blended learning adoption model for higher education," *Education and Information Technologies*, vol. 25, no. 3, pp. 1639–1662, May 2020, doi: 10.1007/s10639-019-10019-z.

[8] A. Balderas, L. De-La-Fuente-Valentin, M. Ortega-Gomez, J. M. Dodero, and D. Burgos, "Learning Management Systems Activity Records for Students' Assessment of Generic Skills," *IEEE Access*, vol. 6, pp. 15958–15968, 2018, doi: 10.1109/ACCESS.2018.2816987.

[9] M. Awad, K. Salameh, and E. L. Leiss, "Evaluating Learning Management System Usage at a Small University," in *Proceedings of the 2019 3rd International Conference on Information System and Data Mining*, New York, NY, USA, Apr. 2019, pp. 98–102. doi: 10.1145/3325917.3325929.

[10] L. M. Regueras, M. J. Verdú, J. D. Castro, and E. Verdú, "Clustering Analysis for Automatic Certification of LMS Strategies in a University Virtual Campus," *IEEE Access*, vol. 7, pp. 137680–137690, 2019, doi: 10.1109/ACCESS.2019.2943212.

[11] H. Henderi, Q. Aini, A. D. Srengini, and A. Khoirunisa, "Rule based expert system for supporting assessment of learning outcomes," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 1.2, pp. 266–271, 2020.

[12] S. Jain, P. Lodhi, O. Mishra, and V. Bajaj, "StuA: An Intelligent Student Assistant," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 2, pp. 17-25, 2018, doi: 10.9781/ijimai.2018.02.008.

[13] S. Hossain, D. Sarma, Fatema-Tuj-Johora, J. Bushra, S. Sen, and M. Taher, "A Belief Rule Based Expert System to Predict Student Performance under Uncertainty," 2019 22nd International Conference on Computer and Information Technology (ICCIT), 2019, pp. 1-6, doi: 10.1109/ICCIT48885.2019.9038564.

[14] G.-J. Hwang, H.-Y. Sung, S.-C. Chang, and X.-C. Huang, "A fuzzy expert system-based adaptive learning approach to improving students' learning performances by considering affective and cognitive factors," *Computers and Education: Artificial Intelligence*, vol. 1, p. 100003, Jan. 2020, doi: 10.1016/j.caeai.2020.100003.

[15] A. Milad *et al.*, "An Educational Web-Based Expert System for Novice Highway Technology in Flexible Pavement Maintenance," *Complexity*, vol. 2021, p. 6669010, Feb. 2021, doi: 10.1155/2021/6669010.

[16] V. M. Ramesh, N. J. Rao, and C. Ramanathan, "Implementation of an Intelligent Tutoring System using Moodle," in *2015 IEEE Frontiers in Education Conference (FIE)*, Oct. 2015, pp. 1–9. doi: 10.1109/FIE.2015.7344313.

[17] A. Q. AlHamad, N. Yaacob, and F. Al-Omari, "Applying JESS rules to personalize Learning Management System (LMS)using online quizzes," in *2012 15th International Conference on Interactive Collaborative Learning (ICL)*, Sep. 2012, pp. 1–4. doi: 10.1109/ICL.2012.6402213.

[18] F. Cervantes-Pérez, J. Navarro-Perales, A. L. Franzoni-Velázquez, and L. de la FuenteValentín, "Bayesian Knowledge Tracing for Navigation through Marzano's Taxonomy," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 6, pp. 234–239, 2021, doi: 10.9781/ijimai.2021.05.006.

[19] M. Constapel, D. Doberstein, H. U. Hoppe, and H. Hellbrück, "IKARion: Enhancing a Learning Platform with Intelligent Feedback to Improve Team Collaboration and Interaction in Small Groups," in *2019 18th International Conference on Information Technology Based Higher Education and Training (ITHET)*, Sep. 2019, pp. 1–10. doi: 10.1109/ITHET46829.2019.8937348.

[20] L. Huang, C.-D. Wang, H.-Y. Chao, J.-H. Lai, and P. S. Yu, "A Score Prediction Approach for Optional Course Recommendation via Cross-User-Domain Collaborative Filtering," *IEEE Access*, vol. 7, pp. 19550–19563, 2019, doi: 10.1109/ACCESS.2019.2897979.

[21] A. Muñoz, J. Lasheras, A. Capel, M. Cantabella, and A. Caballero, "OntoSakai: On the optimization of a Learning Management System using semantics and user profiling," *Expert Systems with Applications*, vol. 42, no. 15, pp. 5995–6007, Sep. 2015, doi: 10.1016/j.eswa.2015.04.019.

[22] S. Sridharan, D. Saravanan, A. K. Srinivasan, and B. Murugan, "Adaptive learning management expert system with evolving knowledge base and enhanced learnability," *Education and Information Technologies*, May 2021, doi: 10.1007/s10639-021-10560-w.

[23] 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, pp. 542–550, 2014.

[24] F. D. de la Pena Esteban, J. A. Lara Torralbo, D. Lizcano Casas, and M. A. Martinez Rey, "Expert system for problem solving in distance university education: The successful case of the subject 'operations management,'" *Expert Systems*, vol. 36, no. 5, p. e12444, Oct. 2019, doi: 10.1111/exsy.12444.

[25] M. Agaoglu, "Predicting Instructor Performance Using Data Mining Techniques in Higher Education," *IEEE Access*, vol. 4, pp. 2379–2387, 2016, doi: 10.1109/ACCESS.2016.2568756.

[26] S. San-Martín, N. Jiménez, P. Rodríguez-Torrico, and I. Piñeiro-Ibarra, "The determinants of teachers' continuance commitment to e-learning in higher education," *Education and Information Technologies*, vol. 25, no. 4, pp. 3205–3225, Jul. 2020, doi: 10.1007/s10639-020-10117-3.

[27] C.-Y. Su, Y.-H. Li, and C.-H. Chen, "Understanding the Behavioural Patterns of University Teachers Toward Using a Learning Management System," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 16, no. 14, Art. no. 14, Jul. 2021.

[28] E. García, C. Romero, S. Ventura, and C. de Castro, "A collaborative educational association rule mining tool," *The Internet and Higher Education*, vol. 14, no. 2, pp. 77–88, Mar. 2011, doi: 10.1016/j.iheduc.2010.07.006.

[29] 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, doi: 10.9781/ijimai.2020.05.006.

[30] P. D. Reddy and A. Mahajan, "Expert System for Generating Teaching Plan Based on Measurable Learning Objectives and Assessment," in *2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT)*, Jul. 2016, pp. 207–208. doi: 10.1109/ICALT.2016.61.

[31] O. Biletska, Y. Biletskiy, H. Li, and R. Vovk, "A semantic approach to expert system for e-Assessment of credentials and competencies," *Expert Systems with Applications*, vol. 37, no. 10, pp. 7003–7014, Oct. 2010, doi: 10.1016/j.eswa.2010.03.018.

[32] J. Fritz, "LMS Course Design As Learning Analytics Variable," in *Proceedings of the 1st learning analytics for curriculum and program quality improvement workshop*, Edinburgh, Scotland, UK, 2016, pp. 15–19. Accessed: Sep. 06, 2021. [Online]. Available: http://ceur-ws.org/Vol-1590/

[33] C. Iwasaki, T. Tanaka, and K. Kubota, "Analysis of Relating the Use of a Learning Management System to Teacher Epistemology and Course Characteristics in Higher Education," 2011, doi: 10.34105/j.kmel.2011.03.032.

[34] J. Whitmer, N. Nuñez, T. Harfield, and D. Forteza, "Patterns in Blackboard Learn tool use: Five Course Design Archetypes." Blackboard, 2016.

[35] T. Harfield, "Analytics for Learn: Using Data Science to Drive Innovation in Higher Education - Blackboard Blog," Apr. 27, 2017. https://blog.blackboard.com/analytics-for-learn-data-science-innovation-higher-education/ (accessed Jan. 26, 2022).

[36] J. Fritz, T. Penniston, M. Sharkey, and J. Whitmer, "Scaling Course Design as a Learning Analytics Variable," in *Blended Learning*, Routledge, 2021.

[37] I. Bennacer, R. Venant, and S. Iksal, "Towards a Self-assessment Tool for Teachers to Improve LMS Mastery Based on Teaching Analytics," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12884 LNCS, pp. 320–325, 2021, doi: 10.1007/978-3-030-86436-1_28.

[38] S. Abu-Naser, M. Barak, and A. Barak, "A Proposed Expert System For Guiding Freshman Students In Selecting A Major In Al-Azhar University, Gaza," Sep. 01, 2008.

[39] E. Mosqueira-Rey, V. Moret-Bonillo, and Á. Fernández-Leal, "An expert system to achieve fuzzy interpretations of validation data," *Expert Systems with Applications*, vol. 35, no. 4, pp. 2089–2106, Nov. 2008, doi: 10.1016/j.eswa.2007.09.045.

[40] E. Verdú, M. J. Verdú, L. M. Regueras, J. P. de Castro, and R. García, "A genetic fuzzy expert system for automatic question classification in a competitive learning environment," *Expert Systems with Applications*, vol. 39, no. 8, pp. 7471–7478, 15 2012, doi: 10.1016/j.eswa.2012.01.115.

[41] A. J. Viera and J. M. Garrett, "Understanding interobserver agreement: the kappa statistic," *Family Medicine*, vol. 37, no. 5, pp. 360–363, May 2005.

[42] Y. Li, "University Teachers' Pedagogical Work with Canvas An exploration of teachers' conceptions, design work and experiences with an LMS," Master Thesis, Universitetet i Oslo, 2019. Accessed: Sep. 09, 2021. [Online]. Available: https://www.duo.uio.no/handle/10852/73114

[43] P. S. Muljana and T. Luo, "Utilizing learning analytics in course design: voices from instructional designers in higher education," *Journal of Computing in Higher Education*, vol. 33, no. 1, pp. 206–234, Apr. 2021, doi: 10.1007/s12528-020-09262-y.

[44] A. Goudarzi, C. Spehr, and S. Herbold, "Expert decision support system for aeroacoustic source type identification using clustering," *The Journal of the Acoustical Society of America*, vol. 151, no. 2, pp. 1259–1276, Feb. 2022, doi: 10.1121/10.0009322.

Luisa M. Regueras

She received her Master's and Ph.D. degrees in telecommunications engineering from University of Valladolid, Spain, in 1998 and 2003, respectively. She is currently an Associate Professor at the Department of Signal Theory, Communications and Telematics Engineering, University of Valladolid. Her research interests include new e-learning technologies, gamification, educational data mining and expert systems.

María Jesús Verdú

She received her Master's and Ph.D. degrees in telecommunications engineering from the University of Valladolid, Spain, in 1996 and 1999, respectively. She is currently an Associate Professor at the Higher Technical School of Telecommunications Engineering, University of Valladolid. She was also Deputy Director for seven years. She has experience in coordinating projects in the fields of new telematic applications for the Information Society and telecommunication networks, especially related to e-learning. Her research interests include new e-learning technologies, gamification, educational data mining and expert systems.

Juan-Pablo de Castro

He received his Master's and Ph.D. degrees in telecommunications engineering from the University of Valladolid, Spain, in 1996 and from the Polytechnic University of Madrid in 2000, respectively. He is currently an Associate Professor at the Department of Signal Theory, Communications and Telematics Engineering, University of Valladolid. He was the Research Director of the Technological Centre for the Development of Telecommunications (CEDETEL) from February 2001 to June 2003. He currently acts as an R&D and technology consultant. His research interests include new e-learning technologies, gamification, educational data mining, expert systems and spatial data infrastructures.

Teaching through Learning Analytics: Predicting Student Learning Profiles in a Physics Course at a Higher Education Institution

Elvira G. Rincon-Flores[1,2,5], Eunice Lopez-Camacho[3], Juanjo Mena[4], Omar Olmos[5]*

[1] Institute for the Future of Education, Tecnologico de Monterrey (Mexico)
[2] School of Humanities and Education, Tecnologico de Monterrey, GRIAL research group, Universidad de Salamanca (Spain)
[3] Independent Researcher (United States)
[4] Education Department, Salamanca University, Institute of Psychology and Educations Kazan Federal University (Russia)
[5] School of Engineering and Science, Tecnologico de Monterrey (Mexico)

Received 24 November 2020 | Accepted 16 October 2021 | Published 25 January 2022

Abstract

Learning Analytics (LA) is increasingly used in Education to set prediction models from artificial intelligence to determine learning profiles. This study aims to determining to what extent K-nearest neighbor and random forest algorithms could become a useful tool for improving the teaching-learning process and reducing academic failure in two Physics courses at the Technological Institute of Monterrey, México ($n = 268$). A quasi-experimental and mixed method approach was conducted. The main results showed significant differences between the first and second term evaluations in the two groups. One of the main findings of the study is that the predictions were not very accurate for each student in the first term evaluation. However, the predictions became more accurate as the algorithm was fed with larger datasets from the second term evaluation. This result indicates how predictive algorithms based on decision trees, can offer a close approximation to the academic performance that will occur in the class, and this information could be use along with the personal impressions coming from the teacher.

Keywords

Adaptive Learning, Educational Innovation, Higher Education, Learning Analytics, Predictive Algorithms.

DOI: 10.9781/ijimai.2022.01.005

I. Introduction

Learning Analytics (LA) is understood as statistical work and computer science applied to educational environments to enhance learning. Educational processes generate much data that can be used to generate actionable insights to innovate the way students learn. LA represents a huge opportunity in higher education for teachers, researchers, and education officials [1]. However, it is not hard to find many educational institutions that have a vast amount of data, but it is not utilized to improve educational processes [2] and [3]. A major limitation is a fact that the data is generated after the courses are completed [4], too late to provide timely feedback to the student and to offer adaptative measures to improve their learning. Also, researchers like Vieira, Parsons & Byrd [5], & Wong [6] claim that currently there is not yet a knowledge field that combines effectively LA and educational theory. In addition, most of what has been developed in this direction are related to online education [7] and [8] and cannot be applied to face-to-face instruction.

The damage caused by the COVID-19 pandemic has spread to most of the countries of the world, not only in the field of health and economics but also to education [9]. While some educational institutions decided to stop their academic activities, others decided to continue online. This represented a significant change for both teachers and students, which has led to rethinking the teaching-learning process [10]. These changes have also affected students emotionally and have urged educational institutions to design strategies to minimize the damage. The Association for Psychological Sciences [11] summarized mental health impacts in children and adults: increased stress levels due to isolation (with consequences on mental and physical health), and anxiety and depression due to overflow of information. This problem takes on various nuances in a world in which inequality prevails. While some educational institutions have the resources to offer online educational environments, others do not have an adequate technological infrastructure to deal with confinement, or rather than following a traditional system schools do not have innovative methodologies to optimally cope with the effects of the pandemic [9] and [10]. Thus, taking as a starting point the ecosystems of educational institutions that have the resources to continue with classes online, it is important to promote motivation to reduce frustration, promote autonomy, flexibility, and frequently give positive feedback [12]. Likewise, it could be valuable for both teachers and students to have a predictive algorithm that allows the teacher

* Corresponding author.
E-mail address: oolmos@tec.mx

to have a general forecast about the performance of the group and thus redouble efforts in those cases with grim predictions. Similarly, the student could have the option of knowing their forecast to better manage their academic period.

Olmos et al. [13] developed an Artificial Intelligence (AI) algorithm based on classification algorithms to forecast the students' academic performance in a Physics engineering course. The forecasting model uses the algorithms K-nearest neighbor and random forest at its core. K-nearest neighbor is a non-parametric algorithm that works by grouping data in similar sets for continuous or categorical prediction [14]. On the other hand, the random forest algorithm uses a group of decision trees, each of them initiated randomly and independently [15]. To begin the process of predicting academic performance, both algorithms are trained with data from a similar sample of students. In 2016, the model training was initially carried out using biometric information such as neuronal frequencies, facial recognition, heart rate, as well as student academic information [13]. Then, to make the predictions of the target group, the algorithm received the students' some set of grades of each of the various activities from the first evaluation period, one by one, making a new prediction with the grades of each activity. In the fourth run, the photographs of each student were used for facial geometric recognition and the final prediction was obtained. In Fig. 1 it can be seen how the predictive algorithm acquires more precision as more data it receives more input.

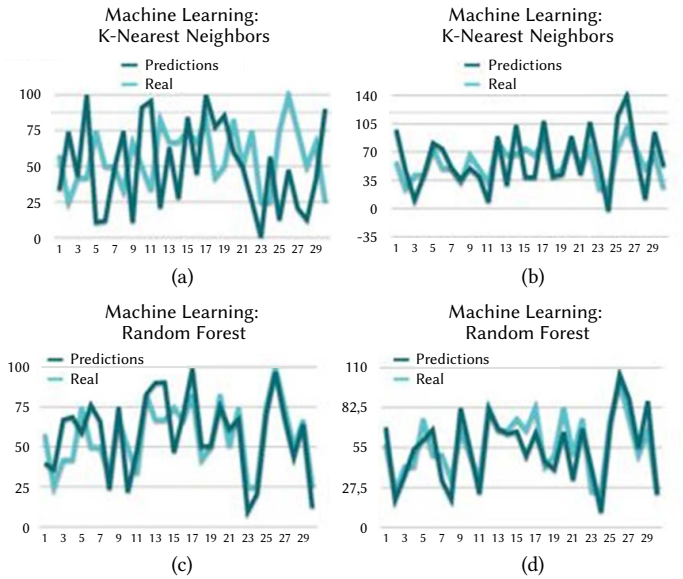


Fig. 1. Algorithm training with a) Quizzes. b) Quizzes + Homework (HW). c) Quizzes + HW + Students surveys to evaluate teachers (ECOA). d) Quizzes + HW + ECOA + Biometrics.

Based on these promising results, we decided to replicate the study with two Physics courses: Physics I (two groups) and Physics II (three groups). These five groups had different instructors. The course consists of three evaluation periods. This study focuses on the first two terms evaluations. The main objective is to determine to what extent the algorithm could become a useful tool for improving the teaching-learning process and reducing academic failure through the timely implementation of preventive adaptive measures.

This research work first presents a theoretical framework that addresses concepts around the artificial intelligence algorithms used. The importance of testing learning models in education based on predictive algorithms serves as a reference in the application of adaptive learning paths in face-to-face learning environments. Then, the methodology is explained and finally, the most relevant results are presented.

II. Theoretical Framework

In the last five years, the use of prediction models that use artificial intelligence to determine profiles has increased dramatically. Companies such as Netflix, Facebook, and Amazon, are examples of this [16] and [17]. This technological development has also been recently spreading to the educational field. The first focuses on the result of student performance, searching through data and automated mechanisms in intelligent Machine Learning algorithms, to draw a line of action that allows the student to strengthen their deficiencies and gradually achieve their performance objectives [18], such as the ALEKS platform, online tutoring, and assessment program, to supplement math instruction [19]. The second way is related to the adaptation of the instructor's actions, intending to generate the actions necessary to adapt to the student's profile and thereby achieve an improvement in their learning [20] and [21]. Adaptive Learning is possible and has become a field of research, particularly in contexts where a vast amount of data can be gathered, and variables can be more controlled. This is the case of Massive, Open, and Online Courses (MOOCs), other online courses, tutorials, computer games, among others [20]. Adaptive Learning has also been tried in hybrid models or Blend Learning. López, Muniesa, and Gimeno [22] designed an adaptive experience on the Moodle platform, in which users were guided through the instructional design activities, adapting at their own pace, the results were positive.

Learning Analytics aims to collect data, analyze it, and generate insights about the students and their contexts to improve their learning, as well as the learning environment [23]. Thanks to advances in the field of computing, interesting opportunities have been created to collect and analyze big datasets [5]. Many algorithms can be applied to the data to discover patterns and make predictions. Random forest is an example of such algorithms [24]. Random Forest uses a combination of predictive variables and a group of decision trees. Each tree provides a prediction based on a random sample of data. The most common prediction becomes the final prediction of the model [15], [25]. Thus, with the help of artificial intelligence, specifically, the use of random decision tree algorithms, it is possible to analyze and identify patterns in populations with a large number of variables [26]. Another algorithm is the K-Near Neighbors which has the purpose of agglutinating the information in similar groups [14].

The combination of algorithms is an innovative model in educational processes since it is presented as a combination of statistical, probabilistic, and forecasting models, in addition to being affordable for most educational institutions.

III. Methods

A. Methods

This study is a result of a research project entitled: "Adaptive based on predictive" carried out at the Tecnologico de Monterrey (Mexico) in a Physics course for undergraduate students. The research embraces a mixed and quasi-experimental methodological approach [27] and [28] with an emphasis on the quantitative data (QUANàqual) to validate and complement the quantitative findings.

A questionnaire of fourteen Likert-type items was handed out to participants. Responses were ranked from 0 (completely disagree) to 5 (completely agree). The research study followed four phases [29]: a.Questionnaire design and validation, b.application and uploading of the online questionnaire, c. statistical data analysis, and d. final report and improvement proposals.

1. Materials

The materials used were the inputs that were used to train the algorithms and to conduct the forecasts. They consisted of teachers' notes, students' photographs, academic records, and the teachers' grades of two terms evaluation.

2. Participants

The sample consisted of 268 graduate students enrolled in Physics I and Physics II courses in engineering. Each course was divided into control and experimental groups. Five instructors (teachers) delivered the instructional content.

3. Tasks and Methods

Initially, the algorithms were trained with the graduate students' grades and self-photographs from previous editions of the Physics I and Physics II courses. In a second phase, the predictions of grades for the first evaluation period were calculated for the experimental groups, using the photographs of the students as the only input. Photography as input was used to recognize facial features so that the algorithm could generate an association matrix, when there is not enough academic information on students n < 5 activities. The information generated from the identification image of each student is used as a reference mark, associating the semantic vector extraction property of the image [13]. At this very moment, the forecasts were released to each teacher so that they applied adaptive measures based on these grades' predictions. The teachers communicated these adaptive measures to the research team through interviews. In a third phase, after the end of the first evaluation period, three forecasts were recalculated for each group, one using only the photography as input, another using photography and the first evaluation period grade, and the third, using only the first evaluation period grade. As the last phase, interviews were done with the students at the end of the course to know their perception of the possibility of knowing the prediction of their grades.

The algorithms used to calculate the forecasts of the first-period grades were K-Nearest Neighbors and Random-Forest. Only the Random-Forest was used to forecast the second evaluation period since it was the one that gave the best result the first time.

To evaluate the precision of the forecasts, 2 error measures were used, and a very elementary reference forecast was constructed for comparison purposes. The error measures used were: Mean Absolute Deviation (MAD) and Average Absolute Percentage Error (MAPE). Equations (1) and (2) show how these measurements are calculated.

$$\left(\frac{1}{n}\right)\sum|Actual - Forecast| \tag{1}$$

$$100 * \left(\frac{1}{n}\right)\sum\frac{|Actual-Forecast|}{Actual} \tag{2}$$

where n is the number of data entries.

The second partial grades were collected from both the control and experimental groups to determine the extent of the adaptive routes in the control groups. Averages are compared with the parametric hypothesis test using the Student's t-statistic.

It is important to mention that, at the beginning of the data collection process, the students signed a letter in which they agreed that their academic photographs be used for this research.

IV. Results

A. Results

The results of the forecasts of the five instructors for the first and second evaluation periods are shown in Fig. 2.

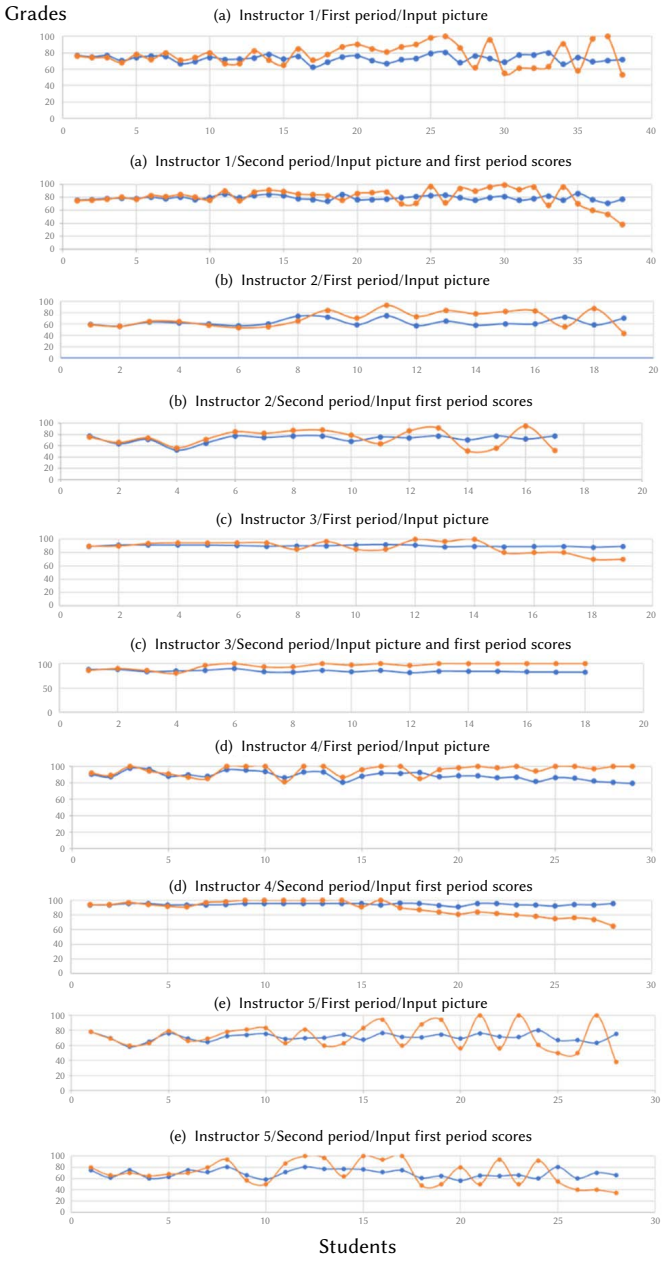


Fig. 2. Predictive and real results of the first and second evaluation period. The Physics I instructors are 1 and 2, the rest are from Physics II (blue= forecasted student performance, orange=actual performance).

Fig. 3 shows the set of adaptative measures implemented by each instructor once they knew the first forecasts for their groups. Each instructor chose what kind of intervention to do for those students with a failing grade forecast. The idea was to prevent the forecast to happen in those cases.

In a second stage, three predictions were calculated for the second evaluation term: one using only photography as input, another using photography and grades from the first evaluation period, and a third using only the grades. Then the best of the three forecasts was selected for each teacher. For instructors 1 and 3 the best forecasts were first period grades and photographs; while for Instructors 2, 4, and 5, the best forecast was generated using only the first-period grades. See Fig. 4. As it might have been expected, grades from the first evaluation period were a good predictor for the second-period grades.

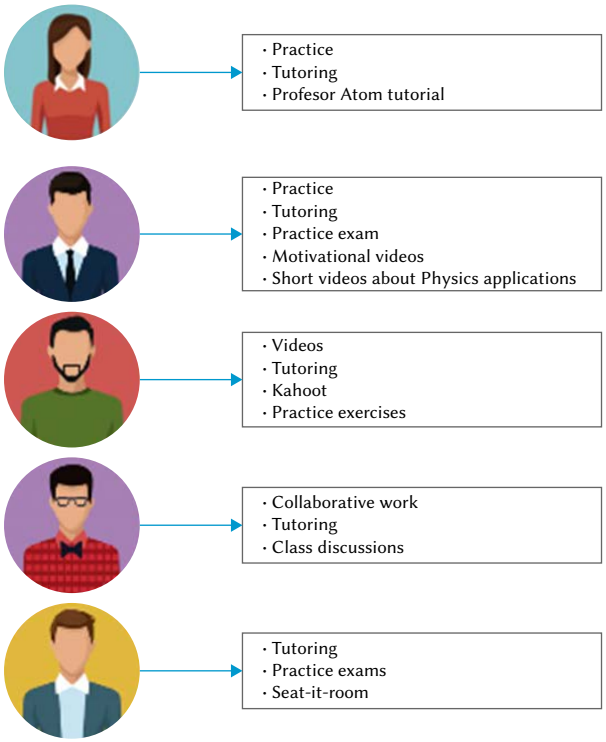


Fig. 3. Summary of interviews with instructors. The Physics I instructors are the first two, the rest are from Physics II.

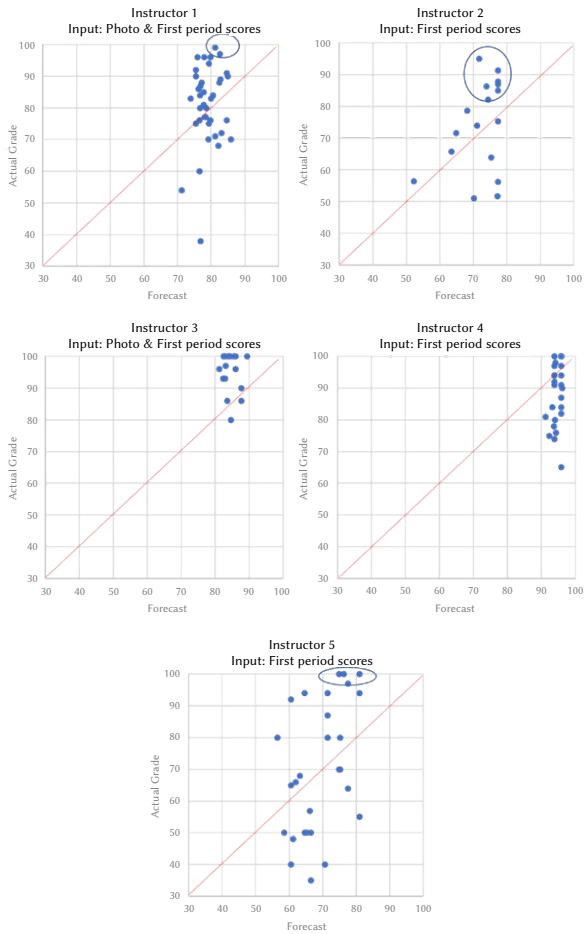


Fig. 4. Predictive and actual students' performance from the second term evaluation using different inputs (Experimental groups). Note that the highest grades in 3 of the groups correspond to some of the highest predictions.

Comparing the first and second term forecasts, the gap observed in the second one is smaller, particularly in the results obtained from instructors 2, 3, and 4. It can be noted that the forecasts generated by the model have a smaller standard deviation than the actual grades because grades vary based on many factors.

To validate the quality of the forecasts generated by the algorithm, a reference forecast was built as a basis for comparison [30]. A reference forecast is also called a naïve forecast and it must be very simple. In our case, the prediction for each student will be simply the average of the forecasts of all the students with the same instructor. This reference forecast can also be called the do-nothing forecast because it does not differentiate between students in the same class.

Fig. 5 and Fig. 6 show the Mean Absolute Percentage Error of the forecast for the second period generated by the algorithm and the reference forecast. Random Forest forecasts had an average error of 16.4% meanwhile the reference forecast average error is 17.8%. As expected, the reference forecast average error is higher for each of the 5 instructors when we use the grades as the only input. However, the difference is very slight (especially for Instructor 3, where the difference of 0.1% is hard to see in Fig. 5).

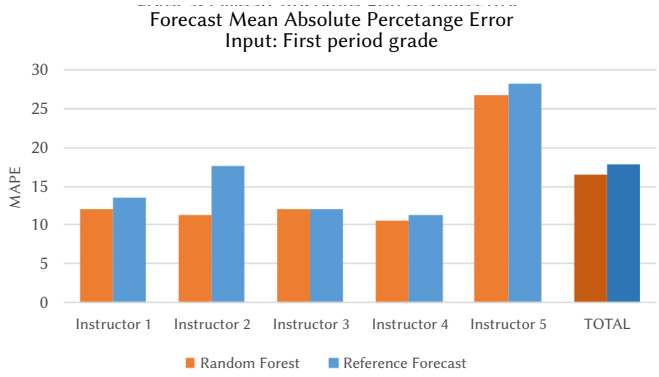


Fig. 5. Validation of the Random Forest forecast compared against a reference forecast. Input: First-period grades.

Surprisingly, when we use the grades and photographs as input (Fig. 6) the difference between both forecasts is less obvious.

Also, second-period grades were harder to predict for instructor 5 class. This may be because this instructor evaluated the first period using some assignments and evaluated the second period with an exam.

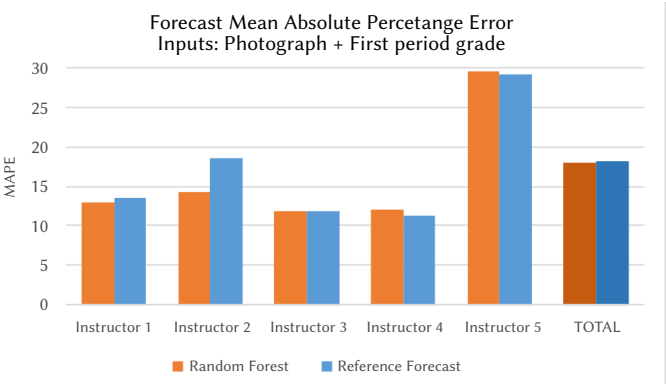


Fig. 6. Validation of the Random Forest forecast compared against a reference forecast. Input: Photograph + First period grades.

Table I presents average results for the first and second evaluation periods of all groups along with the forecast MAD. Note that forecast tends to have a larger MAD in groups when grades have more deviation since it's usually harder to predict values with more dispersion.

TABLE I. Grades From the First and Second Evaluation Periods. The Mean Absolute Deviation of the forecast Produced by the Random Forest Algorithm Is Included

		Period 1			Period 2		
	n	Average	Std Dev	MAD	Average	Std Dev	MAD
Instructor 1 Control	31	78.6	13.1	11.7	80.5	12.2	6.2
Instructor 1 Experimental	38	77.2	12.9	11.9	80.9	12.6	11.1
Instructor 2a Experimental	17	67.8	13.9	11.8	74.1	14.4	11.3
Instructor 3 Control	19	93.9	5.8	5.5	95.4	5.5	11.5
Instructor 3 Experimental	18	88.6	9.3	7.6	95.4	6.2	11.9
Instructor 4 Control	31	95.0	8.4	10.6	90.8	13.3	7.4
Instructor 4 Experimental	28	95.5	5.8	8.3	89.4	10.0	8.1
Instructor 5 Control	58	73.8	15.2	12.8	82.3	19.3	19.2
Instructor 5 Experimental	28	72.3	16.8	13.3	70.6	20.9	16.4

[a] Instructor 2 only taught one group which was considered an experimental group.

From the interviews, information was obtained regarding the adaptive processes applied by each instructor and in each evaluation period, which lasted five weeks. It is important to clarify that given the number of students per group and that the courses are face-to-face it was complex for teachers to take personalized adaptive measures, so they opted for group measures. Instructors 1 and 2 taught Physics I course, based on the predictions calculated before the first and second periods. They decided to provide more practical exercises give more tutoring time. Instructor 2 also showed the students motivational videos and videos about the subject matter. Instructor 3 was in charge of two honors Physics II groups, that is, students who receive more advanced training than those in a regular course. In the first period of the course, he made available a series of videos that help to complement or review the classes, as well as personal tutoring to the experimental group. In the second period, he added immediate response online tests such as Kahoot, questionnaires for review, and more practice exercises. This instructor is willing to use the Predictive Algorithm as a didactic tool. Instructor 4 had two regular courses of Physics II. This instructor decided that the experimental group would be the group that seemed to be lagging behind. Instructor 4 applied collaborative work and offered tutoring in the first period to the experimental group. In the second period, he changed the collaborative work to in-class discussions as plenary sessions. This instructor does not plan on using the Predictive Algorithm as a didactic tool. Lastly, Instructor 5 also taught two regular Physics II courses. This instructor did not distinguish between the control group and the experimental group, and in the first period offered tutoring and practice exams. In the second period, in addition to providing the same options as in the first evaluation period, he changed the sitting arrangement of the problematic students. Instructor 3 is willing to use the Predictive Algorithm as a didactic tool.

In order to analyze the effectiveness of the adaptive routes applied by each instructor, except for instructor 2 who led only one course, the means of the grades of the second period of both the control group and the experimental group were calculated to determine if the differences were significant. The conclusions are shown in Table II and Fig. 5.

TABLE II. Averages of the Second Evaluation Period for the Control and Experimental Groups of the Three Instructors [a]

Instructor	**Average of the control / experimental group**	$P(T \leq t)$ **one-tail**	**Conclusion**
1	80.55 / 80.89	0.454	There is no statistical evidence that the average of both groups corresponds to different populations.
3	95.42 / 95.39	0.493	There is no statistical evidence that the average of both groups corresponds to different populations.
4	90.84 / 89.43	0.325	There is no statistical evidence that the average of both groups corresponds to different populations.
5	82.26 / 70.57	**0.006**	Statistical evidence that the experimental group obtained a lower average than the students in the control group.

[a] Instructor 2 only taught one group, which was considered an experimental group. Therefore, no comparison control vs experimental is done for instructor 2.

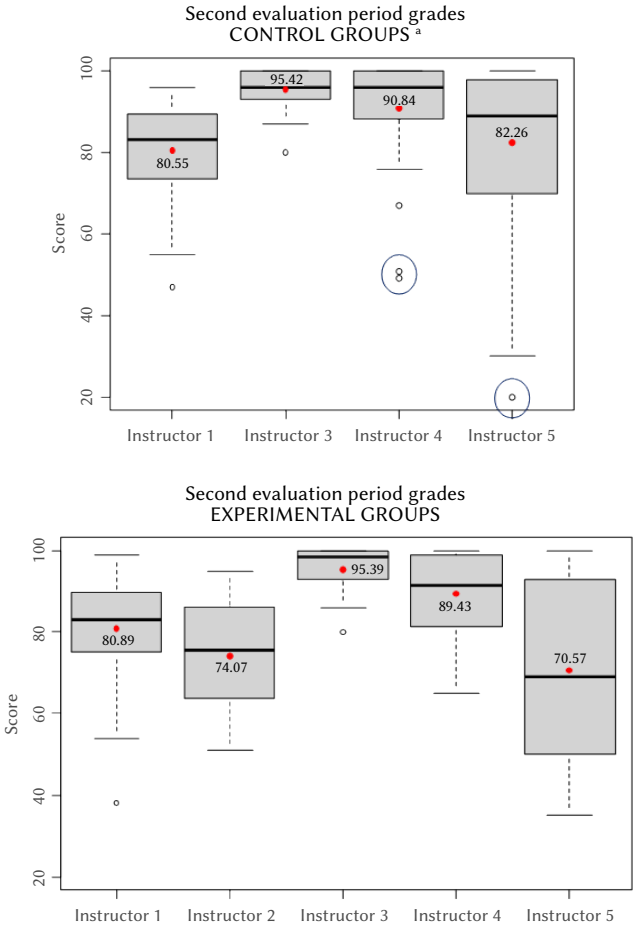
Another result observed pointed out a significant difference between the undergraduate student's mark averages in both groups led by Instructor 5. The interpretation can be explained by comparing with other instructor's groups. For instance, Instructor 3 had honors groups (advanced students) while Instructor 4 opted that the experimental group would be the group more behind, which implies that he managed to raise the academic level of the group. Regarding Instructor 5, dissonance was noted between the forecast and what happened in the classroom, which arouses suspicion about inconsistencies in the teaching-learning process.

Fig. 5 presents a boxplot for the second-period grades for each group. It shows that the variation among grades given by each instructor is similar in their control and experimental groups (the box height is similar). The groups of Instructor 5 are those that show greater dispersion. The four control groups show atypical observations in the range of low scores. That is, some students obtained atypically low marks concerning the rest of the group. In the experimental groups, only two atypical points are observed in the groups of Instructors 1 and 3. For instructors 4 and 5, although the average of the experimental group is not better than that of the control group, atypically low grades are observed only in the control groups (indicated in Fig. 7 with a circle).

Although it is true that the experimental group was expected to have a better result in statistical terms, it is worth considering external factors that could affect the empirical study: e.g., , class schedule, students' group characteristics, teacher's pedagogical style, etc. Furthermore, it is not possible to know what the performance of the experimental group would have been if the teacher had not taken adaptive actions based on the predictions. On the other hand, an improvement is noted in the groups of instructors 1, 2, and 3 when compared to the qualifications of the first and second periods of the experimental groups (Table I).

We wanted to know the students' opinions about the potential use of the algorithm as a learning tool. At the end of the program, and after learning of what this research was about, students voluntarily answered a questionnaire to let us know their opinion about the scenario when they know their own prediction at the beginning of the semester and what they would do if they were giving their grade forecast. A total of

45 students submitted their answers. Questionnaires results show that around 80% of the undergraduates agreed and strongly agreed with the grades they got for assignments, quizzes, and mid-term exams. In other words, their grades aligned with what they expected.



[a] Instructor 2 only taught one group, which was considered experimental group.

Fig. 7. Comparative boxplots. Grades from the second term in both experimental and control groups. The horizontal line inside the boxes represents the median. The little red circle inside the boxes represents the mean and it is labeled with its value.

Also, when they were asked about the chance of using AI to predict their final course exams, they declared that they would like to know the prediction of their academic performance in their courses. See Fig. 8.

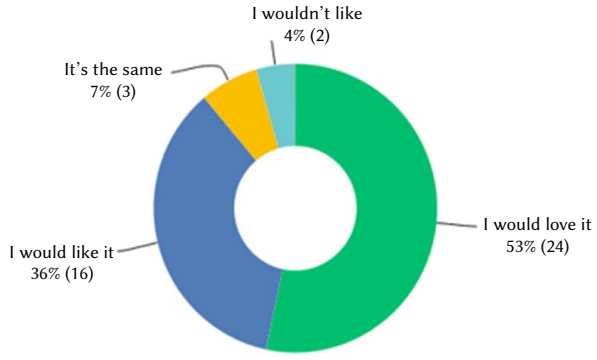


Fig. 8. Undergraduates' opinions about knowing in advance their academic performance prediction by using AI.

In general terms, 53% of the undergraduates would love to use it and 36% would like it. Few students (4%) disagreed with the idea of using this technology to know their performance prediction beforehand.

Another question was about the reason why the undergraduates would like to use the predictive algorithm as part of their academic tools. The answers were grouped into five categories, which are shown in Fig. 9.

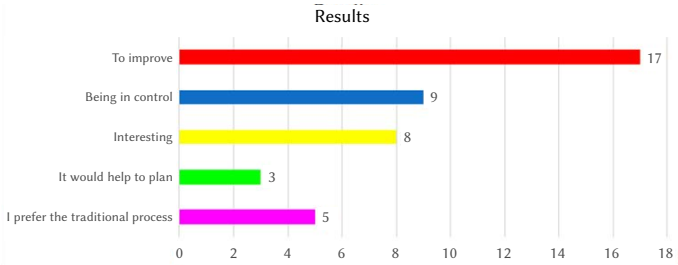


Fig. 9. Undergraduates' opinions about why they would use or not AI to know their forecasted performance.

It is interesting how most students are eager about using predictive algorithms to improve or have greater control of their grades (red color bar). At the same time, other students showed some resistance (purple color bar). Some outstanding opinions of the students were the following:

"It could allow planning and detecting on time what are your weak topics, and you will know what grade you need to stay away from the risk of losing a scholarship" (Student 1).

"It would be interesting, but it could also be a little terrifying" (Student 2).

V. Discussion

Predictions become more accurate as of the algorithm trains with a larger amount of data. This is observed in Fig. 1, where the predictions improve as the inputs were made. This matches with what [31] found: the final academic performance of the students could be predicted with greater precision when a third of the semester had already elapsed. So, decision tree-based AI processes can offer more accurate predictive algorithms as they are fed by a greater amount of data of the current course.

Predictive algorithms, based on decision trees, can offer a close approximation to the academic performance that will occur in the classroom in general. This can be seen in Fig. 2 and 4, where the trend between the first and second evaluations is similar. [4] warn in their study that computer-assisted assessments are the best predictor for the detection of low-performing students. This shows that a blend learning format should be adopted in the face-to-face courses using a technological platform. The computerized resource should be selected consensually to homogenize the activities and evaluations.

Although the predictions were not accurate in terms of students' evaluations, they provide an overview of the group's performance. In this sense, the instructor can take group adaptive measures to preventively improve student performance. On the other hand, Adaptive Learning can be complex because the role of the instructor as a mediator is difficult to perform. [32] found in their study that mediators can influence the dynamics of social learning and adaptive learning. Therefore, the use of predictive algorithms can be a didactic tool that favors adaptive learning, offering the possibility of improving teaching practice and student performance.

In this study, both teachers and students agree to use this type of technology as a tool, both to improve teaching practice and to

improve academic performance, respectively. However, although today's technology can support vast data processing, unlike the 1970s when AI began [19] and [33], there are serious limitations in its use due to the data protection law and ethical aspects [34] that strongly advise not to use facial recognition, for instance, as an input for these algorithms [35].

In this regard, it turned out significant how three of the five predictions were more accurate than using the students' academic records alone. Therefore, colleges and universities could facilitate access to a greater number of data and thereby improve the predictive power of algorithms by covering greater possibilities and reducing the margin of error. In other words, if the algorithm is trained based on students' academic history such as recorded homework, quizzes, and exams, it could be possible to accurately predict that those with a good track record will get good grades. However, it might happen otherwise, as factors such as teacher pedagogical style, students' study habits or personal circumstances may influence the real results. For this reason, feeding the algorithm with a greater amount of data encompasses a greater number of possibilities for future teaching. For example, it will ease teachers' work if they forego initial evaluations of students' aptitude or spot students with difficulties in the course of the teaching program. Or, if the teachers can foresee any group performance trend, then grouping students with low, regular, and high levels could be optimal, as proposed in the work of Villagrá-Arnedo et al. [36].

VI. Conclusion

Based on the research results, the algorithm delivered a forecast of the group performance in general. Therefore, the algorithm can be a valuable resource for the instructor to design and implement adaptative measures. We expect that the forecast will be more precise on final grades.

At the same time, we expect that the individual forecast will be more accurate when the algorithm uses a larger number of variables. In face-to-face courses, there is a legal limitation about the use of personal data, which is an obstacle to getting better data timely.

Given that the forecasts of three out of the five instructors had a smaller margin of error using only the student's academic information, the possibility of making predictions without using facial recognition as the input remains open, eliminating concerns about ethical questions.

Acknowledgment

The authors would like to acknowledge the financial support by Experimentation and Impact Measurement of Tecnologico de Monterrey, Mexico, in this research. We also thank the instructors who participated in the project: Dr. Santa E. Tejeda Torres, MSc. Jorge A. Lomas Treviño, Dr. Alfonso Serrano Heredia y Dr. Héctor J. Medel Covaxin. The work on this study by Juanjo Mena was performed according to the Russian Government Program of Competitive Growth of Kazan Federal University

References

[1] B. Daniel, "Big data and analytics in higher education: Opportunities and challenges," in *British Journal of Educational Technology*, vol. 46, no. 5, pp. 904-920, 2015, doi: 10.1111/bjet.12230

[2] F. J. Garcia-Penalvo *et al.*, "Opening learning management systems to personal learning environments," *Journal of Universal Computer Science*, vol. 17, no. 9, pp. 1222–1240, 2011.

[3] W. Greller and H. Drachsler, "Translating learning into numbers: A generic framework for learning analytics," *Educational Technology Society*, vol. 15, no. 3, pp. 42–57, 2012.

[4] D. T. Tempelaar, B. Rienties, and B. Giesbers, "In search for the most informative data for feedback generation: Learning analytics in a data-rich context," *Computer and Human Behavior*, vol. 47, pp. 157–167, 2015, doi: 10.1016/j.chb.2014.05.038

[5] C. Vieira, P. Parsons, and V. Byrd, "Visual learning analytics of educational data: A systematic literature review and research agenda," *Computer and Education*, vol. 122, no. March, pp. 119–135, 2018, doi: 10.1016/j.compedu.2018.03.018

[6] B. T. M. Wong, "Learning analytics in higher education: an analysis of case studies," *Asian Association of Open Universities Journal*, vol. 12, no. 1, pp. 21–40, 2017, doi: 10.1108/aaouj-01-2017-0009

[7] A. Martínez-Monés et al., "Achievements and challenges in learning analytics in Spain: The view of SNOLA," *RIED. Revista Iberoamericana de Educación a Distancia*, vol. 23, no. 2, pp-187-212, 2020. doi: 10.5944/ried.23.2.26541.

[8] Á. Fidalgo-Blanco, M. L. Sein-Echaluce, F. J. García-Peñalvo and M. Á. Conde-González, "Using Learning Analytics to improve teamwork assessment," *Computers in Human Behavior*, vol. 47, pp. 149-156, 2015. doi:10.1016/j.chb.2014.11.050.

[9] R. F. Arnove, "Imagining what education can be post-COVID-19," *Prospects*, vol. 49, no, 1-2, pp. 43-46, 2020. https://doi.org/10.1007/s11125-020-09474-1

[10] S. J. Daniel, "Education and the COVID-19 pandemic," *Prospects*, vol. 49, no, 1-2, pp. 91-96, 2020, doi:10.1007/s11125-020-09464-3

[11] APA [American Psychological Association]. Human behavior in the time of COVID-19: Learning from psychological science, 2020. https://www.psych ologi calsc ience .org/obser ver/human -behav ior-in-thetime-of-covid -19

[12] M. S. C. Thomas and C. Rogers, "Education, the science of learning, and the COVID-19 crisis," *Prospects*, vol. 49, no, 1-2, pp. 87-90, 2020, doi:10.1007/s11125-020-09468-z

[13] O. Olmos, M. Hernández, E. Avilés, I. Treviño,. "Optimal Paths for academic performance supported by artificial intelligence," *Conference Proceedings of the 6th International Conference on Educational Innovation*, CIIE 2018. Monterrey, Mexico, 2018.

[14] G. Chirici *et al.*, "A meta-analysis and review of the literature on the k-Nearest Neighbors technique for forestry applications that use remotely sensed data," *Remote Sensing of Environment, Elsevier Inc.*, vol. 176. pp. 282–294, 01-Apr-2016, doi: 10.1016/j.rse.2016.02.001

[15] V. K. Ayyadevara, "Random Forest" in: *Pro Machine Learning Algorithms*," Berkeley, CA: Apress, 2018, pp 105-116, doi: 10.1007/978-1-4842-3564-5_5

[16] Y. Koren, "The bellkor solution to the netflix grand prize," *Netflix Prize Doc.*, no. August, pp. 1–10, 2009.

[17] T. Havens, "Netflix. In From Networks to Netflix" *Routledge*, pp. 321-331, 2019, doi: 10.4324/9781315658643-30

[18] Y. Chen, X. Li, J. Liu, and Z. Ying, "Recommendation System for Adaptive Learning," *Applied Psychological Measurement*, vol. 42, no. 1, pp. 24–41, 2018, doi: 10.1177/0146621617697959

[19] B. T. Smith, "How adaptive learning really works," *Tech & Learning*, vol. 37, no.3, pp. 20-26, 2015.

[20] D. Miliband, "Choice and Voice in Personalised Learning," *Pesonalising Education*, pp. 9–19, 2006, doi:10.1787/9789264036604-2-en

[21] F. J. Gallego-Durán, R. Molina-Carmona, and F. Llorens-Largo, "Measuring the difficulty of activities for adaptive learning," *Universal Access Information Society*, vol. 17, no. 2, pp. 335–348, 2018, doi: 10.1007/s10209-017-0552-x

[22] D. L. López, F. V. Muniesa, and Á. V. Gimeno, "Aprendizaje adaptativo en moodle : tres casos prácticos Adaptive learning in moodle : three practical cases," *Education in The Knowledge Society (EKS)*, vol. 16, pp. 1–12, 2015.

[23] D. Gašević, S. Dawson, and G. Siemens, "Aprendizaje adaptativo en moodle : tres casos prácticos Adaptive learning in moodle : three practical cases," *TechTrends*, vol. 59, no. 1, 2015.

[24] Y. Chen, X. Li, J. Liu, and Z. Ying, "Recommendation System for Adaptive Learning," *Applied Psychological Measurement*, vol. 42, no. 1, pp. 24–41, 2018, doi: 10.1177/0146621617697959

[25] M. Belgiu and L. Dragu, "Random forest in remote sensing: A review of applications and future directions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, 2016, doi: 10.1016/j.isprsjprs.2016.01.011

[26] K. Chen, L. Fine, & B. Huberman, "Predicting the Future". *Information Systems Frontiers*, vol. 5, no. 1, pp. 47-61, 2003.

[27] M. Castañer, O. Camerino, M. T. Anguera, "Métodos mixtos en la investigación de las ciencias de la actividad física y el deporte". *Apuntes Educación Física y Deportes* 112, 31-36, 2013.

[28] J. W. Creswell, "A concise introduction to mixed methods research", Thousand Oaks :SAGE, 2015.

[29] S. Olmos, J. Mena, E. Torrecilla & A. Iglesias.Olmos. "Improving graduate students learning through the use of Moodle," *Educational Research and Reviews*, vol. 10, no. 5, pp. 604-614, 2015.

[30] M. Gilliland, "The Business Forecasting Deal," New Jersey, USA: John Wiley and sons, 2010.

[31] O. H. T. Lu, A. Y. Q. Huang, J.C.H. Huang, A. J. Q., Lin, H. Ogata, S. J. H. Yang, "Applying learning analytics for the early prediction of students' academic performance in blended learning," *Educational Technology and Society*, vol. 21, no. 2, pp. 220–232, 2018.doi: 10.2307/26388400

[32] B. Szijarto and J. B. Cousins, "Making Space for Adaptive Learning," *American Journal of Evaliation*, pp. 1–17, 2018, doi: 10.1177/1098214018781506

[33] EduTrends, "Aprendizaje y evaluación adaptativos (adaptive learning and evaluation)," *Observatorio de Innovación Educativa del Tecnológico de Monterrey*, no. Julio, 2014.

[34] F. J. García-Peñalvo, "Learning Analytics as a Breakthrough in Educational Improvement," in Radical Solutions and Learning Analytics: *Personalised Learning and Teaching Through Big Data*, D. Burgos, Ed. Lecture Notes in Educational Technology, pp. 1-15, Singapore: Springer Singapore, 2020. doi: 10.1007/978-981-15-4526-9_1.

[35] S. Agarwal and D. P. Mukherjee, "Facial expression recognition through adaptive learning of local motion descriptor," *Multimedia Tools and Applications*, vol. 76, no. 1, pp. 1073–1099, 2017, doi: 10.1007/s11042-015-3103-6

[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. doi: 10.9781/ijimai.2020.05.006.

Elvira G. Rincon-Flores

Elvira G. Rincon-Flores holds a PhD in Education Sciences from the University of Salamanca, Cum Laude thesis. She is a research professor and experimental and impact measurement scientist of the Institute for the Future of Education at the Tecnologico de Monterrey. Currently, she belongs to the National System of Researchers (SNI), to the research groups: GRIAL and GIIE, the University of Salamanca, and Tecnológico de Monterrey, respectively. She has been the leader of two research projects: Predictive Algorithm based on Artificial Intelligence for learning, and Educational Spaces of the Tec 21 model. It also collaborates with the University of Lima in the development of a dynamic platform for Gamification. Her lines of research are: Educational innovation, educational technology, and evaluation of didactic strategies for learning.

Eunice López-Camacho

Eunice López-Camacho holds a PhD in Information Technology and Communications (2012) from the Tecnologico de Monterrey, with a dissertation about producing hyper-heuristics for solving the two-dimensional irregular bin packing problem. Eunice's academic experience is mostly about teaching Probability, Statistics, and Algorithm Analysis. She has a multidisciplinary range of research interests, including statistical applications in education and manufacturing industry, optimization, hyper-heuristics, and evolutionary computation. Her public profile is available at Research Gate (https://www.researchgate.net/profile/Eunice_Lopez5).

Juanjo Mena

Juanjo Mena (PhD) is Associate Professor at the department of Education in the University of Salamanca, Spain. He is the treasurer of the ISATT and research collaborator in Kazan Federal University, Russia. His work in this study was performed according to the Russian Government Program of Competitive Growth of Kazan Federal University. His research interests: mentoring, practicum, Teacher Education and ICT.

Omar Olmos

Omar Olmos is a full-time professor at the Tecnológico de Monterrey. Holds a PhD in nonlinear Physics, actually, he is an academic director of the North Region science department. He has oriented his academic work to the development of effective educational technology for science learning, as well as new models for the evaluation and development of progressive competencies in science courses. He has given workshops and conferences at national and international universities in Mexico, the United States, Spain, Peru, Colombia, and Chile. He is a consultant on issues of educational innovation, sustainable business development, change management and technological transformation models supported by technology. He has developed and supported Artificial Intelligence models for both academic forecasting and data patterns in different domines.

Balance Your Work-Life: Personal Interactive Web-Interface

Soumi Majumder[1]*, Soumalya Chowdhury[2], Nilanjan Dey[2], K. C. Santosh[3]

[1] Department of Business administration in Vidyasagar University, West Bengal (India)
2 Department of Computer Science and Engineering, JIS University, West Bengal (India)
3 KC's PAMI Research Lab - Computer Science, University of South Dakota (USA)

Received 16 April 2021 | Accepted 14 July 2021 | Published 31 August 2021

Abstract

The term work-life balance can be described as a path to manage stresses and burnouts in the workplace. In this Covid-19 pandemic, work-from-home practice includes both personal and professional spaces as employees, more often, stay digitally connected. As a result, personal life hardly can be separated, which will potentially create imbalanced life, which creates problems regarding physical and mental health of the employees. In such unprecedented situations, we are required to maintain and/or integrate balanced work-life. A balanced work-life gives employees a stress-free environment to work and improves employees' mental and physical health conditions and relationships. In this study, we focus on maintaining a proper work-life balance through a monitoring tool, the 'Wheel of Life.' Considering the drastic changes in work culture (due to Covid-19, for example), we introduce an interactive interface based on 'Wheel of life' concept. Our interface helps tune various important factors, such as business, creative, social, love and life purpose, and provides multiple recommendations. The purpose of the study is to assist web users to balance their work-life, improve psychological well-being and quality of life in this unforeseen situation.

Keywords

User Interface, Interactive Interface, Wheel of Life, Work-Life Balance.

DOI: 10.9781/ijimai.2021.08.016

I. Introduction

IN the vast field of human resource management, an important term and the event is work-life balance. It has a strong influence on the lives of all the workers who are engaged in large size or medium-size, or miniature size business organisations. As a plank in the Women's Liberation Movement, the term "work-life balance" is first coined in the U.K. in the '80s. The workers usually worked for 14 to 16 hours a day on an average and six days a week [1]. The long working hours adversely impact social and health paradigms of the workers. In this way, the concept of work-life balance gets importance to reform and re-establishments.

The meaning of work-life balance is a level of prioritisation between personal and professional activities in an individual's life and how an individual can manage his personal and professional life. In the modern management system of the organisation, work-life balance is a very open practice to discuss, and it is a topical issue due to the enhancement of technology. The man-machine interface becomes large and complex [2]. Previously the work from home was difficult or taking the assignment from the organisation and continuing at home was not a widespread practice. There was a vivid distinguish between one's personal life and professional life too. The cloud-based software, different types of mobile technology, usage of smartphones, and the proliferation of the internet have made it more accessible. Though it is an advantage for the business houses, on the other hand, it generates a high pressure towards the human assets and makes their life blur. A common feature of poor work-life balance is stress. In this current time there has been an increase in health issues resulting from physical and mental stress created due to an upsetting disparity between expectation and yield from employees. Employees are the organisation's critical assets; making them healthy and happy, the business house can go a long path smoothly in this competitive market. It is the responsibility of the employers to provide them with a stress-free environment to make the work-life balance in a true sense.

To capture the situation of the Covid-19 pandemic, we can focus on the present scenario of work-life balance and where we can find the various consequences of the pandemic to the work-life of the people [3]. The new trend of working from home means working more at home. Another disturbing trend that has been incorporated due to this situation is called 'bossware.' It is controversial software used by companies to monitor employees under the interpretation of enhancement of productivity. Its collapses the boundaries of work-life balance. It also generates fears that they are surveillance by employers all the time. The result is for employees, working harder and longer.

The family life or personal life is getting hamper in this way. It is also observed that breaks from work, coffee, or lunch invitation from a colleague or friend vanish. Employers' expectation is going beyond the level and gradually setting the norm of work 24/7 [4]. The true meaning of work-life balance has no significance, and it is not justified as well in the context of HR practices. Companies need to establish clear work policies and a suitable work environment that will encourage staff to take breaks and give some time to the family also. Several studies have

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

discovered that long working hours are not suitable for an individual's health. It generates a negative impact on cardiovascular and mental health. Flexible working hours and schedules provide positive effects on health and well-being. Due to poor work-life balance, employees tend to sleep less. As per research, it is said that a 'normal' sleeping schedule should be an average of 7 to 9 hours per night. This adverse effect includes a higher risk of stroke, coronary heart disease, and mental disorders like anxiety and mental depression. It is also shown by a research study that individuals who work 55 hours or more than that per week have a higher risk of stroke than those people who work for standard hours. Sleep deprivation is very much linked with cardiovascular disease and myocardial infarction and stroke-causing death [5]. Due to two major factors, sleep deprivation and prolonged working hours, the health-related quality of work-life is getting poor day by day. In this pandemic time, the need for work-life balance is more important than ever.

Covid-19 pandemic creates an unpredictable and under pressure work environment among the employees [6]. Boundary theory states that people create and maintain physical, temporal, and psychological boundaries around themselves to simplify their operations in the surrounding world [7]. Forming this limit allows employees to minimize interference between work and non-work life. According to the boundary theory, assuming that when working from home during the lockdown period, employees may find it difficult to create and maintain time limits, physical and psychological, so they may encounter some difficulties in maintaining work-life-balance. During the period of isolation at home, the possibility that work life interferes with family life, or family life interferes with work life, or both, is undeniable [8]. The result is blurring lines between work and personal activities because of the new trend to work from home more. A balanced work-life concept gives employees a stress-free environment to work; it improves their mental health and better physical health too, it makes better relationships, it helps to increase the level of employee engagement at work, supporting making people innovative and creative. Natural balance creates workforce more productive, maintains the bridges of happiness fulfillments and finally, work-life balance gives the birth of successful people in life [9].

There is a need of providing the employees with a platform for self-assessment and regulate the factors (known as areas of life) to improve work-life balance. As per our knowledge concern, no similar work has been reported in the literature. The primary motivation of this study is to provide the employees with an interactive platform for real-time self-assessment and tuning area of life (e.g., business life, creative life, social life, love life, etc.) by studying a visual tool, namely: "spiral-web" wheel (a particular type of wheel of life).

In section II, we have discussed the related work. Section III is a detailed description of the wheel of life. Section IV reports the interactive web interface design and implementation for work-life balance, followed by a discussion section (section V). Finally, section VI is the conclusion of this study.

II. Literature Review

Work-life balance is defined and described as a mechanism to manage the increasing amount of stress in the workplace. The typical symptoms of stress have adverse effects on employees' quality of work-life that mainly focus on mental abilities, physical well-being of employees, and their behavior pattern. The stress can come into employees' minds due to continual change of work environment, bullying, lack of challenges, continuous interruption, and a non-effective internal communication system [10]. The researchers have revealed how to manage the problems of stress by implementing simple techniques called the 'Wheel of Life'. This indicates an imbalance life of one individual at present time and points them to address for better management. It mainly focuses on movement and changes in one's life about the above-mentioned factors (work, family, health, friends, love spirit, and wealth) and ultimately generates the wheel or circle that divides into different segments of life. It has been observed that during the 20th century, work demand was going very high, and the personal life of the people was getting hampered. Employers understood and acknowledged the necessity of work-life balance programs, to maintain a healthy balance between one's professional and personal life [11]. The study investigated and analyzed the emergence of work-life balance and established a macro-level model on the work-life balance phenomenon. This facilitated the movement from unbalanced life to balanced life at the organizational level during the 20th century. In the shadow of human resource management and organizational behavior, one of the most important fields under research is work-life balance. It is a phenomenon where personal (spiritual development, family & friends, leisure, pleasure) and work-life (professionalism, career, ambition, esteem needs) both are included. The need for work-life balance depends on the generation, culture, place, individual perception, attitude as well. Research has revealed the conflicts of perceptions of workplace people on their unbalanced life and find out the recommendations to set up a balanced work- life. Maslow's hierarchy model, Vroom's expectancy theory, and Wheel of life strategy had been implemented in this work. This study concluded to focus on the familiarized and transparent management policies. This helps to create changes in the way of better motivation of workplace people to maintain a balanced life in the business organisation. Wheel of life visualization tool has been widely used in various work-life balance studies in the recent past [12], [13]. Work-life balance is the way to solve the problem of increasing pressure in the workplace when people try to balance various factors in the work/life environment, including: family; friends; health; and spirit/ self [14]. A study [15] reported that the stress can be overcome by applying a simple tool called the "wheel of life", which can point out imbalances in an individual's current life and point out ways to solve these problems. The wheel of life is a familiar concept in many religions and spiritual cultures. It represents the constant movement and change in life. Work-life balance is a broad concept that includes adequate prioritization between "work" (career and aspirations) on the one hand and "life" on the other (happiness, leisure, fanaticism, and spiritual development). Needs are also different from generation, culture, place, and personal perception. In [16] authors investigated the views of the generation on work-life balance and, at the same time, organize how to implement work-life balance, identify conflicts of opinion and find possible suggestions to help you maintain a better work-life balance. Recent studies show that the Covid-19 pandemic has significant effects on quality of work-life of the people. The tremendous unbalanced life has been created due to work-from-home practice. People are getting stressed, and the personal life is getting worst. According to the World Health Organization, the novel coronavirus disease has brought the entire world into a relative standstill framework. Due to the absence of effective vaccination, the precautionary measures from the disease are quarantined, closing all workplaces including business houses and schools, spatial distancing, etc. In this situation, the Indian government also takes the actions of 'lockdown' and self-isolation policies. Only a few areas like purchasing for essential items, visit on health issues and other essential work purposes, had some fewer restrictions. The employees had to do work with engagement in virtual mode too much. Work from home means work a lot from home, the result is poor and unbalanced work-life with psychological distress [17]. Another research has been reported on the investigation of the effects of social support and work-life balance on burnout of employees in the context of the Covid-19 pandemic. Researchers took the samples from different sectors like education, IT, health, retailing, tourism, service,

and logistics. The study discovered the relation between burnout, social support, and work-life balance. Findings depicted that social support affects work-life balance and partial effect on burnout. Due to the changing pattern of work-life in the pandemic situation, there is no significant change in male and female employees' state of mind [18], [19]. There is also an adverse impact created by the Covid-19 pandemic and prolongs lockdown to an individual's life in the form of unhappiness, depression, frustration, boredom, fatigue, and other negative emotional symptoms. While work-life balance is found unbalanced, then happiness acts as a protective factor that can help make a better-balanced life [20]. The quality of work-life and work-life balance are entering into uncertainty due to the uncertain nature of the situation. People need to be happy and cherish every sphere of their life to live long with good health. Therefore, their unbalanced life should be balanced and maintained [21]. The literature depicts a need for a real-time self-assessment visual tool for the Employees to understand their current state and in which area (areas of life) they need to improve for better and balanced livelihood.

III. The Wheel of Life

The wheel of life is the perfect tool to monitor an individual's journey towards happiness and success. By using this tool, one can reflect and gain some meaningful insight into the balance of his life [22]. This tool also provides satisfaction in one's life and its different areas. In our study wheel of life, it is focused on goal setting and coaching of the individuals. The purpose is to spend his or her present time, and how much one can be satisfied in different categories of life and work [23]. The wheel of life mainly includes the "Pie" style and the "Spider Web" style. In "Spider Web", style scores are noted on the actual lines for each category (not across the segment). The wheel of life is segmented into different areas. The areas are business life, life purpose, love life, social life, and creative life. Each of the main areas of life is divided into two subcategories as shown in Fig. 1, namely: money & finance, career & work, growth & learning, growth & learning, partners & love, family & friends, environment and community, health & fitness and fun and recreation.

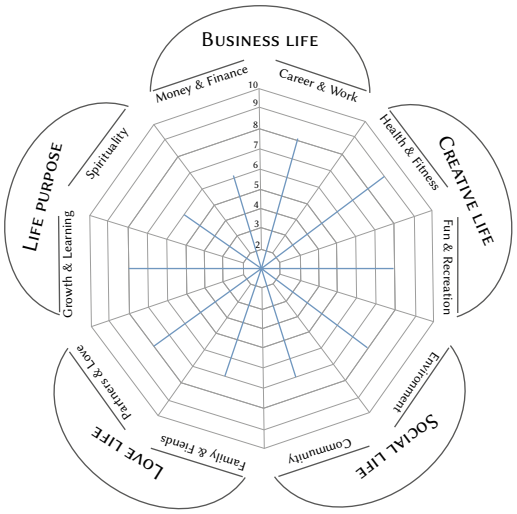


Fig. 1. The wheel of life.

In-wheel of life employees satisfaction levels for each area of life can be scored out of 10, where 1 signifies least satisfaction level and 10 depicts the most satisfactory level. There is always a chance of bias input from the employee, which might affect the overall performance of the proposed model. The user must keep in mind these points before self-assessment: justification of low or high score, an idea about the ideal score for each life area, targeted values (want to achieve) after a month, after three months, after six months, and after a year, etc. Parents need to prioritize work-life balance to lower the level of family conflict, increase job and life satisfaction, reduce stress levels and overall productivity. Most people live complicated lives taking multiple actions in various directions every day. Those who simplify their lives are much happier [24]. Wheel of life is a visual tool to balance work-life and living a happier life.

IV. Interactive Web-interface Tuning for Work-life Balance

The need for interactive web interface design and implementation has been discussed in the previous section. In this section, we will discuss the model design, and implementation of the same.

A. Design and Implementation

As discussed in the earlier section, the wheel of life has various areas of life which can be tuned to balance work-life. Fig. 2. shows the steps of the designing process of the proposed interactive web interface.

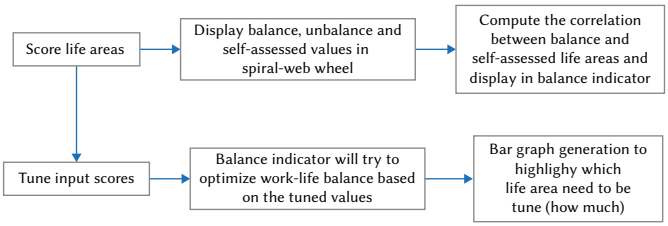


Fig. 2. Interactive web-interface design for living area tuning and work-life balancing.

In this work, an interactive web interface is developed (https://workbalancedlife.web.app/ Last Access Date: 06.07.2021) using JavaScript as a backend language and HTML and CSS as a front-end coding language. In our study, users can provide inputs in 10 points Likert scale for each area of life (Fig. 3) (1 = least satisfied, 10 = most satisfied). The values will reflect (in blue color) in the "spiral web" wheel (Fig. 4). In-wheel of life, red colour shows unbalanced life and green color depicts balanced life where the values are predefined in a previous study [25]. Correlation between green color and blue color is computed to find out closeness between self-assessed work-life and balanced work-life. Based on the correlation value, the balance indicator will shift and try to optimize work-life balance.

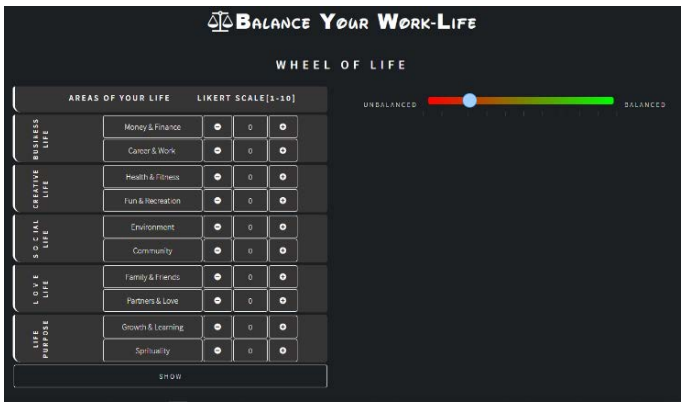


Fig. 3. Areas of life on the Likert scale.

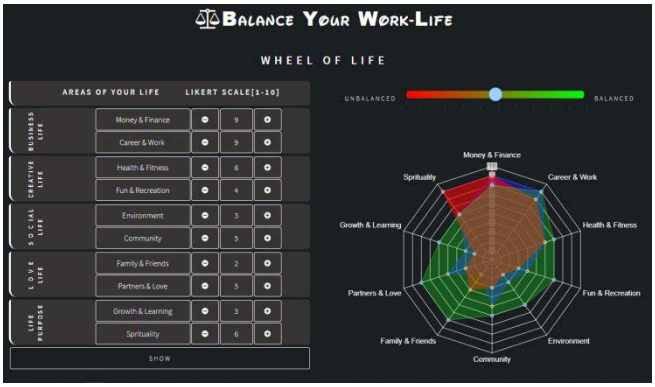


Fig. 4. Wheel of life.

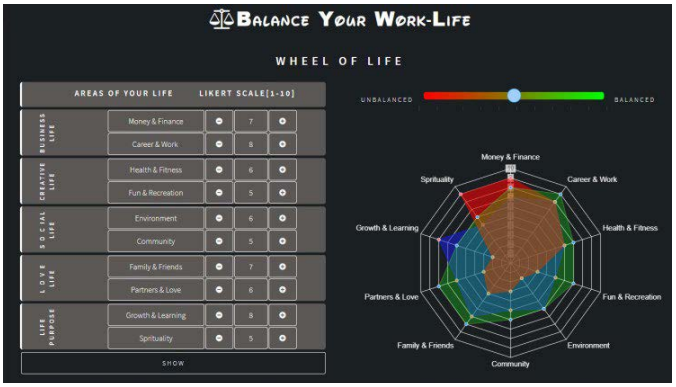


Fig. 6. Wheel of life (Case study).

Fig. 5 shows that the input values by the user will automatically propagate in the tuning table. The user can tune those values, and the balance indicator can be tuned accordingly. Tuned indicators can be compared with ground truth values. Our prime objective is to tune those life areas in such a way so that balance work-life can be achieved. Fig. 5 also shows that depending upon the tuned values (an increase or decrease of the original inputs), a bar graph is reported where the red color indicates which area of life needs to give less focus, whereas Green indicates highly focused.

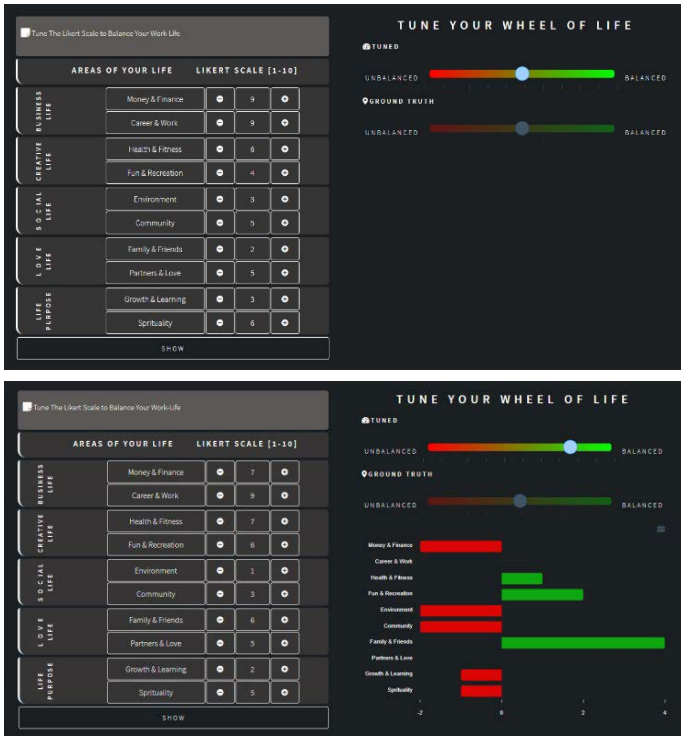
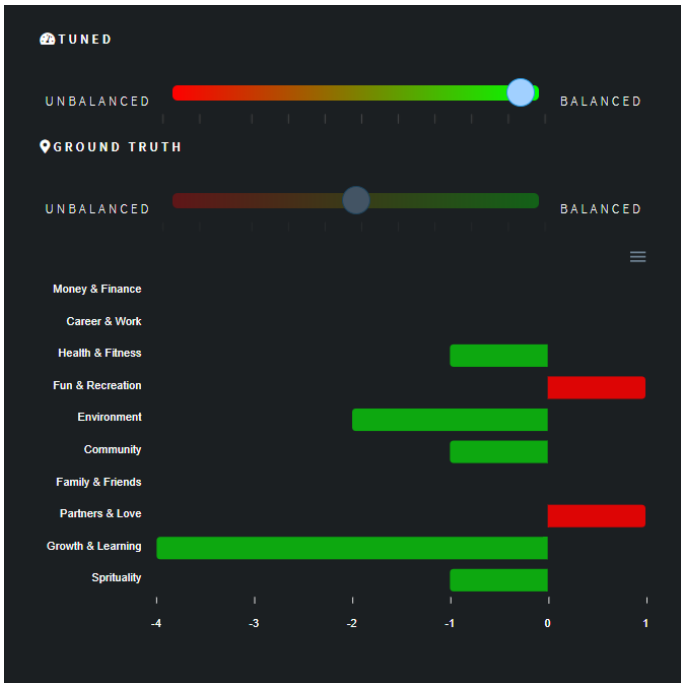


Fig. 5. Tuned value and bar graph.



As we already mentioned, the employee needs to provide their honest self-assessment as input to obtain the correct outcome.

For a given set of initial input values (areas of life) (Fig. 6.), there is always a possibility that the recommendation based upon the tuned values is not satisfactory. In that case, the user can tune multiple times, keeping the input values as-is amongst several recommendations (as shown in Fig. 7). Users can opt for any of the outcomes (bar graphs), which looks like a feasible solution for their work-life balance. The current system allows the user to select the best recommendation (as per choice) amongst several alternatives to obtain a similar outcome (balancing).

Fig. 7. Multiple bar graphs based on tuned values (same inputs).

B. Result Analysis

We have studied the work-life balancing of 120 employees of a small-scale company in Kolkata, India. Amongst 120 employees, 17 employees' work-life is not balanced (<= 50%-marked in red color), average balancing is 40.28% (Fig. 8.) which changed to 90.48% after tuning the areas of life.

Our experimental study shows that (Fig. 9), for business life, 70 employees required positive tuning whereas four required negative tuning. It is interesting to note that 44 employees did not require any tuning in the initial input value, for career & work employees required positive tuning (93) or no change (24) (in the bar graph, green color indicates positive tuning and red signifies negative tuning) (Table I). Sixty-three people required no tuning in creative life, whereas 32 required positive tuning and 25 negative tunings. For fun & recreation, 54 people required no tuning, whereas 50 required positive tuning and 15 negative tunings. In both social life community and environment cases, the majority required either no tune or negative tuning. In Table I majorities' show that attributes namely: love life, family & friends and partners & love, needs improvement. The reason may be the excessive workloads and extended working hours, work-life imbalance during this pandemic time.

Fig. 8. Unbalance-balanced life.

Fig. 9. Areas of life tuning.

Interestingly, most of the employee required negative tuning of growth and learning and spiritual life. Growth is nothing but changing our thinking habits, emotion control, the approach of handling life, etc. It is about change by learning new things. It is most important to retain in the job market in this crisis time rather than learning, growth, and personal development. Maybe this is one reason why most of the employee's growth and learning factors need to be negatively tuned.

Again, as mentioned earlier, employees can change their mindset and always retry for alternative recommendations that might affect the overall results. In this case, we have only considered each employee's very first attempt to change the average work-life balance from 40.28% to 90.48(Fig. 8.).

TABLE I. Tuning Table

Area of Life	No Change required [0]	Negative tuning required [<0] [red]	Positive tuning required [>0] [Green]
Business Life			
Career & Work	24	1	93
Money & Finance	44	4	70
Creative Life			
Health & Fitness	63	25	32
Fun & Recreation	54	15	50
Social Life			
Environment	50	57	12
Community	48	48	23
Love Life			
Family & Friends	28	4	86
Partners & Love	23	6	89
Life Purpose			
Growth & Learning	39	64	16
Spirituality	57	45	17

V. Discussion

Our experimental results clearly show that most employees (Fig. 7) required work-life balance in this crisis period. Areas of life are highly significant to improve the overall quality of work-life as well as personal life. Unbalance in life leads to family conflicts and creates problems in personal relationships besides mental stress and health issues [26]. Poor work-life balance decreases overall productivity, creativity, and engagement level and increases the turnover rate in the workplace. Working from home has erased the boundaries between the time dedicated to working and the time dedicated to oneself in this pandemic time [27]. This changing trend of remote working provides a high level of stress on employees. Employees require continuous mental health support from their organisations. Otherwise, employers will face the consequences of revenue generation, reputation, recognition for a long-term basis, which is not a good sign for any business organisation [28]. Satisfaction and good functioning at work and at home, with a minimum of role conflict [29] is work life -balance. It is always important to study the impact of work life-balance for psychological wellbeing of employees [30], [31]. It is well-established that there are high needs to design an interactive interface that will help the employees balance their work-life.

Our interface will help tune various life areas (business life, creative life, social life, love life, and life purpose) and provide multiple recommendations. Users can choose any one of those recommendations and improve the living areas accordingly for work-life balancing [32]. Our current system is a web-based application. In the future, we plan to design a mobile application. Push notification, convenient interface based on user habits, offline user mode, alert/alarm, etc., can also be incorporated to make the application more robust and user friendly. As a future work of the study, we can also work to deploy decision rules generated by AI algorithms. Always there exists a thin line of defence relying on the honesty of the users who score their areas of life truthfully. The overall performance and efficacy very much depend upon the same. Our platform only recommends which areas of life users need to improve and how much but never suggests which activity needs to be performed to achieve the goal. As a future scope of the study user experience can be studied [33], [34].

VI. Conclusion

In today's rapidly changing modern work environment, time pressure seems to be increasing and new technologies allow work to be done anytime, anywhere. These are just two factors that make it increasingly difficult for workers to integrate work and family life. A balanced work-life gives employees a stress-free environment to work. Improvement of employee's mental and physical health condition, relationship, engagement at work, innovation, and creativity is highly dependent on balanced work-life. Happy employees are efficient and productive, and their level of engagement is very high despite having any uncertainty in a business environment. People should enjoy while working, and it is their responsibility to spend some quality time with personal life and make it successful too. Our study might assist the users to balance their work-life, improve psychological well-being and quality of life in this unforeseen situation.

References

[1] F. Yoshimura and T. Suzuki, "Calcium-stimulated adenosine triphosphatase in the microsomal fraction of tooth germ from porcine fetus," Biochim Biophys Acta, vol. 410, no. 1, pp. 167–177, Nov. 1975.

[2] Harris, H., "Global careers: Work-life issues and the adjustment of women international managers". Journal of Management development, 2004.

[3] P. Naithani, "Overview of Work-Life Balance Discourse and Its Relevance in Current Economic Scenario," ASS, vol. 6, no. 6, p. p148, May 2010.

[4] B. Thomason, and H. Williams, "What will work-life balance look like after the pandemic", Harvard Business Review, pp. 1-4,2020.

[5] K. P. Amin, M. D. Griffiths, and D. D. Dsouza, "Online Gaming During the COVID-19 Pandemic in India: Strategies for Work-Life Balance," Int J Ment Health Addiction, Jul. 2020, Accessed: Jul. 15, 2021.

[6] A. Putri and A. Amran, "Employees' Work-Life Balance Reviewed From Work From Home Aspect During COVID-19 Pandemic," International Journal of Management Science and Information Technology, vol. 1, no. 1, p. 30, Jan. 2021.

[7] L. Qiu and J. Fan, "Family boundary characteristics, work–family conflict and life satisfaction: A moderated mediation model," International Journal of Psychology, vol. 50, no.5, pp. 336-344, 2015.

[8] T. D. Allen, E. Cho, and L. L. Meier, "Work–Family Boundary Dynamics," Annu. Rev. Organ. Psychol. Organ. Behav., vol. 1, no. 1, pp. 99–121, Mar. 2014.

[9] V. H. Patil, N. Dey, P. N. Mahalle, M. Shafi Pathan, and Vinod. V. Kimbahune, Eds., Proceeding of First Doctoral Symposium on Natural Computing Research: DSNCR 2020, vol. 169. Singapore: Springer Singapore, 2021.

[10] W. M. A. Wan Mohd Yunus, S. K. Z. Badri, S. A. Panatik, and F. Mukhtar, "The Unprecedented Movement Control Order (Lockdown) and Factors Associated With the Negative Emotional Symptoms, Happiness, and Work-Life Balance of Malaysian University Students During the Coronavirus Disease (COVID-19) Pandemic," Front. Psychiatry, vol. 11, p. 566221, Feb. 2021.

[11] T. Tuğsal, "The Mediator Role Of Social Support Amid Work-Life Balance And Burnout Of Employees' In The Context Of Coronavirus Pandemic Precautions And Social Isolation," Beykent Üniversitesi Sosyal Bilimler Dergisi, Jul. 2020.

[12] S. C. Wong and A. Ko, "Exploratory study of understanding hotel employees' perception on work–life balance issues," International Journal of Hospitality Management, vol. 28, no. 2, pp. 195–203, Jun. 2009.

[13] S. P. Schwartz et al., "Work-life balance behaviours cluster in work settings and relate to burnout and safety culture: a cross-sectional survey analysis," BMJ Qual Saf, vol. 28, no. 2, pp. 142–150, Feb. 2019.
[14] U. Byrne, "Wheel of Life: Effective steps for stress management," Business Information Review, vol. 22, no. 2, pp. 123–130, Jun. 2005.
[15] U. Byrne, "Work-life balance: Why are we talking about it at all?," Business Information Review, vol. 22, no. 1, pp. 53–59, Mar. 2005.
[16] H. R. Gamage, T. Sailikitha, J. Karamchandani, K. Gowda, and X. X. Tong, Three generations and their work life balance: are we balancing work and life or adjusting life for work?, 2014.
[17] H. R. D. Putranti, S. Suparmi, and A. Susilo, "Work Life Balance (WLB) Complexity and Performance of Employees during Covid-19 Pandemic," Arthatama, vol. 4, no. 1, pp. 56-68, 2020.
[18] A. E. Halaris, K. T. Belendiuk, and D. X. Freedman, "Antidepressant drugs affect dopamine uptake," Biochem Pharmacol, vol. 24, no. 20, pp. 1896–1897, Oct. 1975.
[19] B. A. Sethi, A. Sethi, S. Ali, and H. S. Aamir, "Impact of Coronavirus disease (COVID-19) pandemic on health professionals," Pak J Med Sci, vol. 36, no. COVID19-S4, May 2020.
[20] U. Byrne, "Wheel of Life: Effective steps for stress management," Business Information Review, vol. 22, no. 2, pp. 123–130, Jun. 2005.
[21] J. Greenberg, "Comprehensive stress management," McGraw-Hill Education, 2012.
[22] S. Schieman, P. J. Badawy, M. A. Milkie, and A. Bierman, "Work-Life Conflict During the COVID-19 Pandemic," Socius, vol. 7, p. 237802312098285, Jan. 2021.
[23] T. R. Anderson and T. A. Slotkin, "Maturation of the adrenal medulla--IV. Effects of morphine," Biochem Pharmacol, vol. 24, no. 16, pp. 1469–1474, Aug. 1975.
[24] A. B. Evans et al., "Sport in the face of the COVID-19 pandemic: towards an agenda for research in the sociology of sport," European Journal for Sport and Society, pp. 1–11, May 2020.
[25] A. Deshpande, P. Salunke, and T. Joshi, "Work life balance in phase of pandemic", Bi-lingual Int Res J, vol. 10, no. 38, pp. 229-240, 2020.
[26] J. A. Phillips, "Work–Life Fit During A Pandemic," Workplace Health Saf, vol. 68, no. 10, pp. 502–503, Oct. 2020.
[27] https://medium.com/@erictaussig/balancing-the-wheel-of-life-as-a-working-parent-d4b0c261b084 (last access 4/6/21)
[28] S. Majumder and D. Biswas, "COVID-19: impact on quality of work life in real estate sector," Qual Quant, Mar. 2021.
[29] S. C. Clark, "Work Cultures and Work/Family Balance," Journal of Vocational Behavior, vol. 58, no. 3, pp. 348–365, Jun. 2001.
[30] I. John, N. K. Anthony, and D. Y Bakari, "Impact of Work Life Balance on the Psychological Wellbeing of Employees in the University of Cape Coast.," JPBS, vol. 8, no. 1, 2020.
[31] N. Dey, R. Mishra, S. J. Fong, K. C. Santosh, S. Tan, and R. G. Crespo, "COVID-19: Psychological and Psychosocial Impact, Fear, and Passion," Digit. Gov.: Res. Pract., vol. 2, no. 1, pp. 1–4, Jan. 2021.
[32] S. Schieman and A. Narisada, "A less objectionable greed? Work-life conflict and unjust pay during a pandemic," Research in Social Stratification and Mobility, vol. 71, p. 100564, Feb. 2021.
[33] M. Schrepp, R. Otten, K. Blum, and J. Thomaschewski, "What Causes the Dependency between Perceived Aesthetics and Perceived Usability?," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 6, p. 78-85, 2021.
[34] M. Schrepp and J. Thomaschewski, "Design and Validation of a Framework for the Creation of User Experience Questionnaires," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 5, no. 7, p. 88-95, 2019.

Soumi Majumder

Soumi Majumder, completed her PGDM, specialization in Human Resource Management from the Institute of Business Management, Jadavpur University in 2012. She is a PhD. student of the Department of Business administration in Vidyasagar University, Midnapore, West Bengal, India. Currently, she is working as an Assistant Professor in the Department of Management Science, Sister Nivedita University, Kolkata, India. Previously, she was associated with Techno India College of Technology, NSHM College of Management and Technology, Dinabandhu Andrews Institute of Technology and Management, J D Birla Institute of Science and Commerce, West Bengal State Labor Institute, Siliguri, etc. She published 12 research articles in various international journals and conferences. Her research area includes employee engagement, job satisfaction, leadership, stress management, job burnout, industrial relations, competence learning model, quality of work-life, safety management etc.

Soumalya Chowdhury

Soumalya Chowdhury, is currently doing his B.Tech. In Computer Science and Engineering from JIS University, Kolkata, Inida. His research interests include web designing, machine learning and cyber security.

Nilanjan Dey

Nilanjan Dey, is an Associate Professor in the Department of Computer Science and Engineering, JIS University, Kolkata, India. He is a visiting fellow of the University of Reading, UK. He is an Adjunct Professor of Ton Duc Thang University, Ho Chi Minh City, Vietnam. Previously, he held an honorary position of Visiting Scientist at Global Biomedical Technologies Inc., CA, USA (2012–2015). He was awarded his Ph.D. from Jadavpur University in 2015. He is the Editor-in-Chief of the International Journal of Ambient Computing and Intelligence (IGI Global). He is the Series Co-Editor of Springer Tracts in Nature-Inspired Computing (Springer), Series Co-Editor of Advances in Ubiquitous Sensing Applications for Healthcare (Elsevier). His main research interests include medical imaging, machine learning, computer-aided diagnosis, data mining, etc.

KC Santosh

KC Santosh is the Chair of the Department of Computer Science at the University of South Dakota (USD). Before joining USD, Prof. Santosh worked as a research fellow at the U.S. National Library of Medicine (NLM), National Institutes of Health (NIH). He was a postdoctoral research scientist at the LORIA research centre (with the industrial partner, ITESOFT (France)). He has demonstrated expertise in artificial intelligence, machine learning, pattern recognition, computer vision, image processing and data mining with applications, such as medical imaging informatics, document imaging, biometrics, forensics, and speech analysis. His research projects (of more than $2m) are funded by multiple agencies, such as SDCRGP, State of South Dakota, Department of Education, National Science Foundation, and Asian Office of Aerospace Research and Development. He is the proud recipient of the Cutler Award for Teaching and Research Excellence (USD, 2021), the President's Research Excellence Award (USD, 2019), and the Ignite Award from the U.S. Department of Health and Human Services (2014). For more information, follow: http://kc-santosh.org and https://www.linkedin.com/company/kc-pami/ (research lab).

UX Poker: Estimating the Influence of User Stories on User Experience in Early Stage of Agile Development

Andreas Hinderks[1]*, Dominique Winter[2], Francisco José Domínguez Mayo[1], María José Escalona[1], Jörg Thomaschewski[3]

[1] University of Seville (Spain)
[2] University of Siegen (Germany)
[3] University of Applied Science Emden/Leer (Germany)

Received 6 May 2021 | Accepted 18 March 2022 | Published 28 November 2022

Abstract

Agile methods are used more and more frequently to develop products by reducing development time. Requirements are typically written in user stories or epics. In this paper, a new method called UX Poker is presented. This is a method to estimate the impact of a user story on user experience before development. Thus, there is the opportunity that the product backlog can also be sorted according to the expected UX. To evaluate UX Poker, a case study was conducted with four agile teams. Besides, a workshop followed by a questionnaire was conducted with all four agile teams. The goal of being able to estimate the UX even before development was achieved. Using UX Poker to create another way to sort the product backlog can be considered achieved in this first evaluation. The results show that UX Poker can be implemented in a real- life application. Additionally, during the use of UX Poker, it was found that a shared understanding of UX began. The participants clarified in the team discussion about UX Poker what related to influence the user stories had on UX and what UX meant for their product.

Keywords

Agile, Agile Methods, Usability, User Experience, User Experience Management, UX Management, UX, UX Estimation.

DOI: 10.9781/ijimai.2022.11.007

I. Introduction

Today's users expect to derive a high level of satisfaction while interacting with a product. They also expect to be able to use the product without having to make any major effort to finish their tasks in a quick and efficient manner. Moreover, for a product to succeed, it is important to consider hedonic qualities, that is, the qualities that are not directly target-oriented [1]. It is, therefore no longer sufficient to develop only usable products, they must also inspire the user and address hedonic qualities. In summary, the user wants to have a positive user experience (UX) while interacting with any product or service.

A well-known definition of user experience is given in ISO 9241-210 [2]. Here user experience is defined as 'a person's perceptions and responses that result from the use or anticipated use of a product, system or service'. Therefore, user experience is viewed as a holistic concept that includes all types of emotional, cognitive, or physical reactions to the concrete or even only the assumed usage of a product formed before, during, and after use. In ISO 9241-220 [3] the term human-centred quality has been introduced. Human-centred quality includes user experience, usability, accessibility, and minimizing risks arising from the use.

An additional interpretation defines user experience as a set of distinct quality criteria [1] that includes the classical usability and non-goal directed criteria [4]. Thus, usability is classified as a set of pragmatic factors or qualities, such as efficiency, controllability, or learnability. Non-goal directed criteria are classified as a set of hedonic factors or qualities [4], such as stimulation, novelty, or aesthetics [5]. This definition has the advantage that it splits the general notion of user experience into a number of quality criteria, thereby describing the distinct and relatively well-defined aspects of user experience. This also complies with ISO 9241-220 [3]. One advantage of this definition is that user experience could be measured by using standardized questionnaires such as UEQ+ [6]–[8], SUPR-Q [9], or VisAWI [10]. In addition, a benchmark [11] or KPI [12] can be calculated based on the individual UX factors. The UEQ+ is a modular framework that allows one to combine predefined UX factors to create a concrete UX questionnaire. Currently, the UEQ+ framework contains 20 UX scales, but they can be extended as needed. The construction of the clarity factor can be read as an example [13].

Software development teams use agile methods to develop products or services more and more efficiently. Agile methods (e.g. Scrum [14], Kanban [15], or Extreme Programming (XP) [16]) reduce the time taken to develop a product and make it available on the market [16]. The iterative approach to developing software minimizes the risk of developing software that is not in line with what is needed in the market [17]. The requirements to be developed are collected, evaluated and prioritized in a product backlog [18]. The items with the highest priority were selected for the next development iteration. This also means that the requirements must be prioritized by some method. In agile methodologies, requirements are typically written in user stories or epics.

* Corresponding author.
E-mail address: andreas.hinderks@iwt2.org

This paper, we will present UX Poker, a method to estimate the impact of user stories or an epic on user experience. We will also present the results of a first evaluation study conducted in four different companies.

This paper is structured as follows: Section II briefly summarizes the related work. Section III present the research methodology, including the evaluation study. Section IV outlines the results and key findings of our evaluation study. Section V discusses the meaning of the findings, the limitations of our evaluation study, and the improvements that could be made in it. The paper ends with Section VI, with conclusions and ideas for future work.

II. Background and Related Work

In general, requirements are collected and sorted in agile methods in a product backlog. At least that is what the Scrum Guide [13] requires. Also, ISO 9241-210 [2] and ISO 9241-220 [3] recommends a sorted list of requirements. In all cases, it is not defined which criteria would be used for sorting.

In the literature, there are many papers that investigate the integration of UX Methods and Agile development. The range of methods includes usability engineering, user-centred design (UCD) or human-centred design (HCD) [3], and UX methods in general. [19] conducted a systematic mapping study in 2017. The purpose was to investigate artefacts used in communication between Agile methods and user-centred design. A total of 20 artefacts were identified and examined, such as prototype, user story, scenario, sketch, persona, and card, like the design card or the task-case card. During the development iteration, about 56% of the artefacts were used. The rest were used during the discovery or planning phase.

User stories, prototypes (low and high), sketches and mock-ups are the artefacts with which a UX professional can communicate goals or requirements between developer and stakeholder [19], [20]. These artefacts are usually good at representing both UX and functionality [19], [21]. In practice, the items in the product backlog, mostly written as a user story or epic, are sorted by their importance. A user story is typically described according to the following pattern: “As a [persona], I want [some goal] so that [some reason]”. The goal of this writing style is to present the requirements shortly and understandably. With “persona” the target group of the user story is named, with “some goal” the actual requirement is named and with “some reason” a justification for the user story is named.

In a product backlog the most important user story is at the top of the list, the least important user story, further down. Here there is no clear definition of what is or is not important. There are different methods to determine the importance. Classically, the product owner decides which items are important based on discussions with the stakeholders. But business or marketing requirements can also influence the importance of a product backlog item. Another possibility could be to include the expected user experience in the sorting.

Choma et al. [22] extended or supplemented the grammar of a user story with user experience aspects and usability requirements. New or replaced components of a UserX Story include personas, goals, interactions, contexts, and feedback. Nielsen’s heuristics serve as the acceptance criterion. Expected user experience aspects can be specified as heuristics. Based on these heuristics, the user experience could be estimated by extending and using a suitable method.

Joshi at al. [23] provide a Usability Goals Achievement Metric (UGAM). This metric is calculated by using individual parameters per usability quality (such as learnability, speed of use, and ease of use) weighted to a goal parameter score. This is the goal to be achieved. After each usability evaluation, UX professionals calculate the achieved score based on the values from the usability evaluation. This makes it possible to determine whether the goal has been achieved by comparing the goal with the archived value. If the goal has not been achieved, it is possible to determine where it has not been achieved for each usability quality.

The last two described approaches are not directly based on product backlog items or requirements. Neither approach provides the possibility of estimating the user experience. In the end, both approaches can be used with an appropriate estimation method. Instead of the goal value, an estimated value of the user experience can be specified. The necessary prerequisites for a user experience value to be compared before development are given in both the approaches.

In our view, it is necessary to consider not only usability aspects [23], [24], but also user experience in general. Besides, from our point of view, the agile team should be involved, so that its expertise is also used. In addition, all team members should have the same or similar understanding of UX for their product. Therefore, we have developed UX Poker as a new method, presented in the next section.

III. Research Methodology

In this section, we will describe our approach in detail. Our approach is divided into two main steps:

- Step 1: Method to estimate user experience for a given user story or epic (see A).
- Step 2: First Evaluation of the method from Step 1 by conducting a study with four development teams (see B).

The different steps are explained in greater detail in the next two paragraphs.

A. UX Poker

UX Poker is inspired by Planning Poker [25]. The goal of Planning Poker is to estimate the complexity of a user story or an epic. This estimation is used as a basis for the selection of user stories for the next development iteration. It is a support to fill the next iteration with realizable user stories so that they can be implemented within the iteration. Planning Poker focuses on the technical implementation of the functionality described in the user story. The objective of Planning Poker is to create a consensus about the complexity of a user story. The result of Planning Poker is recorded in a user story and ideally reviewed in a retrospective. The review should result in improvements in the use of Planning Poker. If possible, Planning Poker should result in realistic values of complexity. However, this is an individual and iterative learning process of the Agile team. We applied this idea of Planning Poker to UX Poker as well.

UX Poker is a method that aims to estimate the possible impact of a user story or an epic on the user experience, that is produced at the user’s site. Before prototypes are created, or the actual development begins, the influence of a user story or an epic on the user experience must be determined. In the end, a user story has been evaluated not only in terms of technical implementation but also in terms of the expected UX. Thus, before the actual development starts, the user stories for the next iteration can be explicitly selected based on the expected UX. For example, if the attractiveness of the product is to be increased, user stories that have a significant expeceted influence on the UX factor attractiveness can be specifically selected.

Besides, the team should adopt the user’s perspective through UX Poker. This is to train the team members to look at the development of the product more from the user’s perspective. As a general practice, most of the team members are developers. Therefore, they tend to focus more on the technical implementation of the user stories.

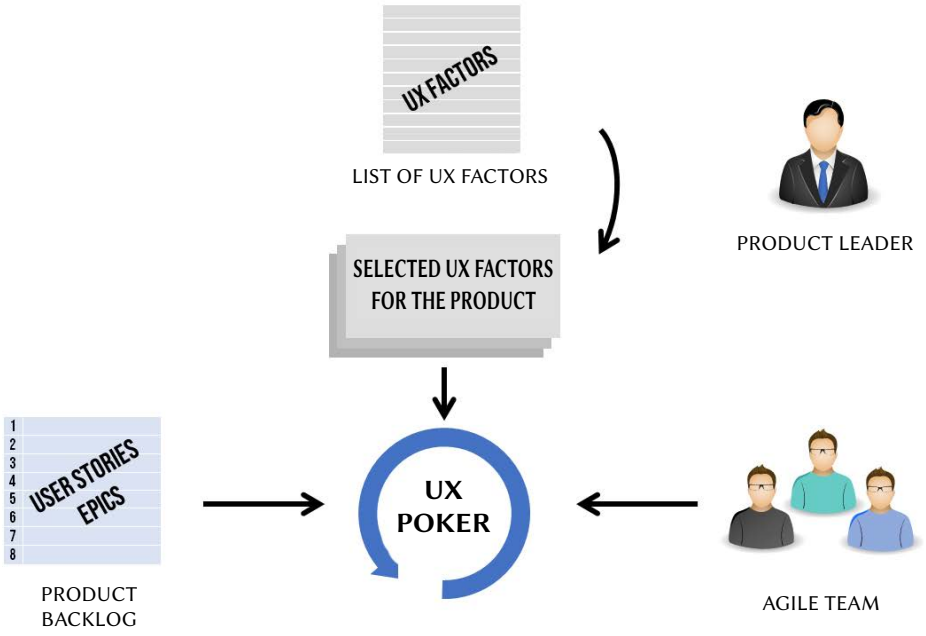


Fig. 1. Procedure of UX Poker with selection of UX factors and UX Poker with product backlog items.

In our opinion, UX Poker makes sense if a product is to be improved in terms of user experience. With UX Poker, a targeted selection of user stories or epics can be made based on the UX estimation before the actual development is made. The goal is to create a basis for a decision about the UX before development starts. If at least a UX estimate of user stories is known before development, the user stories can also be selected specifically.

The procedure of UX Poker is shown in Fig. 1. To use UX Poker, a selection of UX factors for a product is necessary, as described in the next section (see 1). UX Poker as a method is described in sections 2 and 3.

1. Selection of UX Factors

To use UX Poker, a selection of UX factors for a product is necessary. As mentioned in the introduction, user experience can be described using UX factors. This allows to description of specific aspects of the user experience in UX factors. These aspects can be, for example, Efficiency (The user can reach their goals with a minimum time required and minimum physical effort), Quality of Content (Subjective impression if the information provided by the product is up to date, well-prepared and interesting), Attractiveness (Overall impression from the product.), or Trust (The product appears trustworthy to the user). The listed examples of factors are certainly not complete. A good overview is provided by Schrepp and Thomaschewski [26] or Hinderks et al. [27].

UX Poker is based on UX factors to describe aspects of user experience and estimate these aspects for user stories or epics. Instead, the introduction of UX Poker must determine which aspects of the user experience are important for the product. For example, trust is certainly a critical UX factor for banking software, but it plays a secondary role in a computer game.

There are different methods to select the important UX factors for a product from a list of UX factors. For example, the method Ranking (sorting UX factors in a team) or Dot Voting [28] (sorting UX factors by prioritizing). Informal consultation between the product owner and a UX professional can also be carried out. In the end, the method used is not decisive. However, no more than 5-7 factors should be selected, or else meaningful estimation of the factors will no longer be attainable. The recommendation for the number of factors is based on the experience of the authors. If the number of factors is too high, there is a risk that UX Poker will become inefficient and therefore the actual goal will not be achieved.

This list of UX factors can be changed after each iteration. It may well be that after a retrospective, it is recognized that UX factors are missing or do not fit. This list of UX factors can be changed after each iteration. It may well be that after a retrospective, it is recognized that UX factors are missing or do not fit. In this case, the list of UX factors should be revised.

It has been shown in practice that the selection of UX factors is quite unproblematic and hardly needs to be adjusted B.

2. Workflow of UX Poker

To use UX Poker, relevant UX factors must be defined in advance to be handled with UX Poker. The selection of UX factors is described in greater detail in section 1. The UX factors must represent important UX aspects of the product being developed so that UX Poker estimates the correct UX factors. The selection of UX factors is done once and can be changed over time.

UX Poker uses the same idea as Planning Poker in agile methodologies, but with a focus on the possible impact of the user story or epic on the user experience. In UX Poker, the following steps are carried out in a team meeting:

1. The Product Owner or Product Manager presents the user story or epic to the team. The team should understand the goal of the user story or epic.
2. For each selected UX factor, the team estimates the potential impact of it on the user experience. Each team member is asked to rate the possible influence of the user story or epic on user experience using a scale from -2 to +2:

   -2: strong negative impact

   -1: slightly negative impact 0: no impact

   +1: slightly positive impact

   +2: strong positive impact

   It is important that all team members make their evaluations secretly and then disclose them together, as Planning Poker in agile methodologies does.

3. If there are deviations of more than two scale steps, the variations are discussed within the team. The goal is to understand why this deviating assessment has occurred. Afterwards, a new estimation is made, as described in step 2.
4. If there are no or slightly deviations within the team, the average value is recorded in the User Story or Epic.

At the end, for every user story or epic, there is a possible impact for each UX factor.

3. Example of UX Poker

For a better understanding, we give an example of a Twitter app. The UX factors Attractiveness, Quality of Content, Trustworthiness of Content and Trust were selected from the list of UX factors as important UX factors for the Twitter app by the product leader.

User story: As a Twitter user, I want to see if a tweet contains an untrue statement so that I can critically question it.

Ratings and explanation for each UX Factor:

1. Attractiveness: +1. It has a positive impact on Attractiveness, because the new function is new and helpful.
2. Quality of Content: 0. It has no impact on this factor, because it is just a new category for the content and has nothing to do with the content itself.
3. Trustworthiness of Content: +2. It has a very high impact on this factor, because it categorizes the content as far as its trustworthiness if concerned.
4. Trust: +1. Overall, it has a positive impact on trust because of its positive impact on trustworthiness of content.

Overall, this user story has a possible influence on most UX factors. This rating can help the agile team to select this epic if the Twitter app has to improve in terms of Trustworthiness of Content. In addition, UX Poker promotes communication and a common understanding within the agile team.

B. First Evaluation of UX Poker

In a first evaluation, we have used the new method UX Poker in a real-life application with four teams from different companies. On the one hand, the goal was to determine whether the method could be implemented with an agile team. On the other hand, we wanted to identify potential for improvement during and after UX Poker use.

For this reason, we split the evaluation into two parts. First, we conducted a workshop with the agile team, using UX Poker on their user stories or epics. After the workshop was conducted, participants were asked to fill out a questionnaire.

1. Workshop

The workshop, which we conducted with various teams, was organized by us as follows:

- Introduction of the participants.
- Presentation of the UX factors previously selected with the product leader.
- Presentation of the user stories or epics by the Product Owner (see 2).
- Estimation of the UX for the user stories or epics (see 2).

Afterwards, the participants had to fill out the questionnaire, which was concluded with a short retrospective. The workshop should not last longer than 1.5 hours.

2. Construction of the Questionnaire

The questionnaire is intended to determine the subjective assessment by the participants of the usefulness of UX Poker. To this end, we developed a questionnaire with two types of items. On one side there were items with a 7-point Likert scale and the other side featured open-ended questions. The results of the items with the rating scale could be statistically well evaluated. For the open-ended questions, we wanted to get feedback, or opinion on the potential for improvement, from the participants.

The questionnaire contains the following items:

1. With UX Poker we were able to talk in a structured way about the influence of the epic on UX. [Do not agree - agree with a 7-point Likert scale]
2. UX Poker helped me to get a better understanding of the targeted UX for our product. [Do not agree - agree with a 7-point Likert scale]
3. What added value do you see in using UX Poker? [open question]
4. How easy was UX Poker to use? [not easy - very easy with a 7-point Likert scale]
5. Can UX Poker be applied to Epics? [absolutely not - absolutely yes with a 7-point Likert scale]
6. What tips would you have if you recommended UX Poker to others? [open question]
7. What worked well when using UX Poker? [open question]
8. What did not work well when using UX Poker? [open question]

We implemented the questionnaire in an online version using LimeSurvey.

3. Context

The study was conducted in Germany with four agile teams via online video conferencing due to the corona pandemic. It was conducted between October 2020 and January 2021. The agile teams work on different products in different companies. An overview of the products developed by the agile teams is shown below.

- Agile team 1 (7 participants): Internal ordering system in the construction industry for enterprise customers.
- Agile team 2 (7 participants): Soccer Portal App for End Users.
- Agile team 3 (9 participants): Platform for mediation craftsmen and customers.
- Agile team 4 (7 participants): Portal for buying and selling real estate.

A total of 30 (5 females, 23 males, 2 not specified) participants took part in the study. The average age is 36 years (37 for females, 35 for males). Table I shows the distribution of the participants' roles within the teams.

TABLE I. The Role in the Team

Role	Count	Avg. Years of Experience
UX Professional	1	10.0
UX Researcher	1	4.0
UX Designer	3	12.3
Product Owner	5	3.6
Programmer	19	10.0
Other	1	4.0
Sum/Avg	30	8.8

The number of programmer participating in the study is noticeable. However, the distribution of the teams is balanced. At least one product owner and one person responsible for UX are involved in the teams. All teams use Scrum as an agile method.

IV. Results

The individual workshops showed that UX Poker was applicable in practice. This was reflected in the results of the questionnaire.

In the next sections we present the results of the individual items of the questionnaire. For items with rating scale, the corresponding statistical data were presented. For items with open text questions, the answers are summarized and presented accordingly.

A. Q1: With UX Poker We Were Able to Talk in a Structured way About the Influence of the Epic on UX

On average, the subjects answered this question with 'mostly agree' (median 2), as shown in Fig. 2. The small confidence interval and the low standard deviation related to the small number of participants indicate a homogeneous evaluation, despite there being the four different teams.

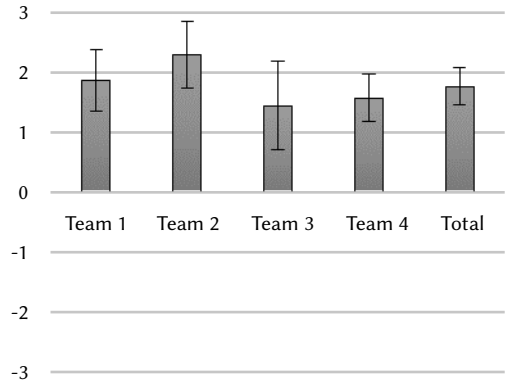


Fig. 2. Result of Question 1 with 95% Confidence Interval as Error Bar.

The total mean value is 1.767 with a variance of 0.737 (Std. Dev. 0.858). The Confidence (95%) is 0.307.

B. Q2: UX Poker Helped Me to Get a Better Understanding of the Targeted UX for Our Product.

Team 1 and Team 3 rated the second question as 'agree' on average, as shown in Fig. 3. Team 2 and Team 4 tended to rate it as 'mostly agree'. On average, the overall result is exactly between 'agree' and 'mostly agree'.

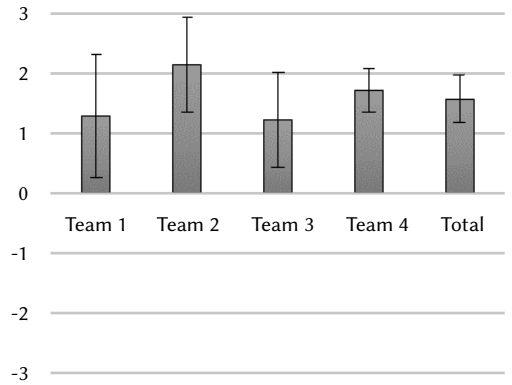


Fig. 3. Result of Question 2 with 95% Confidence Interval as Error Bar.

The total mean value is 1.567 with a variance of 1.220 (Std. Dev. 1.104). The Confidence (95%) is 0.395.

C. Q3: What Added Value Do You See in Using UX Poker?

Almost all participants were optimistic about getting into a conversation and talking explicitly about the user experience. It helped to develop a shared understanding of user experience. It was also noted that UX Poker helped determine which epics had a negative impact on the UX. In Refinement, it was stated that a participant talked, more about the technical implementation. In UX Poker, the user is in the foreground.

It was also stated that UX Poker supported the entire team's ability to participate in the UX process. The UX professional can share their knowledge with the whole team. The authors have similar findings when applying the morphological analysis of 'context of use' [29].

D. Q4: How Easy Was UX Poker to Use?

The result for Question 4 (see Fig. 4), however, ranges from 'agree' (team 4) to 'totally agree' (team 2) if, the confidence interval is taken into account. It is noticeable that the confidence interval in the evaluation of teams 1 and 4 is high compared to that of teams 2 and 3. This is due to the different evaluations provided by the participants and the low number of participants, which is noticeable in the individual result, but is lost in the total (see Total in Fig. 4).

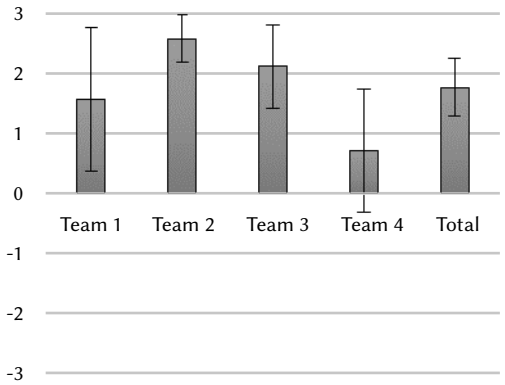


Fig. 4. Result of Question 4 with 95% Confidence Interval as Error Bar.

The total mean value is 1.767 with a variance of 1.771 (Std. Dev. 1.331). The Confidence (95%) is 0.476.

E. Q5: Can UX Poker Be Applied to Epics?

The results for the fifth question are between 'agree' and 'mostly agree', as shown in Fig. 5. As with the previous question, the confidence interval for two teams (teams 1 and 3) is higher than that of the other two teams (teams 2 and 4).

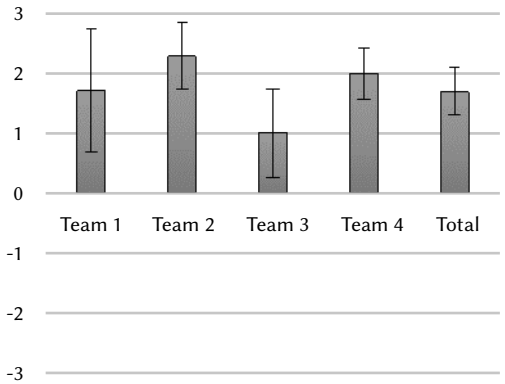


Fig. 5. Result of Question 5 with 95% Confidence Interval as Error Bar.

The total mean value is 1.700 with a variance of 1.183 (Std. Dev. 1.088). The Confidence (95%) is 0.389.

F. Q6: What Tips Would You Have if You Recommended UX Poker to Others?

The main tips provided by participants are to guide the discussion that inevitably takes place during estimation; otherwise, it may run the risk of quickly losing focus. Besides, the presence of fewer participants may unnecessarily prolong the discussion.

For estimation, some test persons indicated that one should use a polling and/or voting tool. This would speed up the process.

However, the participants demanded that the UX vision of the product should be clear in any case. Besides, the UX factors to be estimated should be explained clearly and understandably. Only in this way can UX Poker be used successfully.

G. Q7: What Worked Well When Using UX Poker?

On the positive side, participants stated that UX Poker was very quick to learn and that one could, start right away. It promoted communication within the team about the UX to be achieved. Furthermore, in the constructive discussion, the advantages and disadvantages of the individual Epics could be better assessed from the user's perspective.

Further, the participants stated that the inclusion in the UX processes was positively received. This also promotes visibility of UX processes and resulting activities outside UX Poker.

H. Q8: What Did not Work Well When Using UX Poker?

Some participants were unable to provide any information due to lack of experience. As a point of criticism, the participants stated that a better understanding of the rating scale for UX Poker needed to be created. The scale should be better described so that everyone has the same understanding.

Besides, a common and deeper understanding of the selected UX factors should be created. The participants were sometimes very unsure of what was meant by the UX factors.

V. Discussion

In the results of questions Q1 2, Q2 3, Q4 4 and Q5 5, some differences in the degree of agreement between the individual teams can be seen. We attribute these differences to the different target groups of the teams' products. For example, the target group of Team 1's products is enterprise customers, Team 2's targets are private end customers, Team 3's are craftsmen and private end customers, and Team 4's targets are real estate marketers and private end customers. In addition, the maturity level of the individual teams is different, which would certainly influence the results. However, the tendency is that all results are in the same range when measured against the confidence interval. In the application of UX Poker, new insights have been gained on the use of the method. On the one hand it is the use of the method which is under the consideration of Personas [29]. On the other hand, there is the realization that UX Poker allows a different perspective on epics. We will discuss both points in greater detail in the next sections.

A. The Usage of Personas

In the workshops, a question was repeatedly asked as to which user the UX should be estimated for. It was sometimes not clear to the participants what type of users they should put themselves into the shoes of. The goal of UX Poker is to estimate the UX that will later be created for the user.

For this reason, it makes sense to introduce personas [30] as a prerequisite for UX Poker. Equipped with a clear picture of the personas, UX Poker participants can evaluate the UX from the perspective of these personas. This requirement also coincides with the 'UserX Story' method of Choma et al. [22]. Personas are also a component of this method.

In order to integrate the persona deeply into the development process, it is recommended that persona-driven user stories be used [31]. The user story makes it immediately clear which persona is being addressed.

B. The User Perspective

During the workshops, it became apparent that with UX Poker a different discussion about the implementation of the Epics took place vis-a-vis the exercise in Refinement. Since the participants put themselves into the role of the user, the Epics were analyzed differently. Therefore, things that did not stand out during the Refinement surfaced in the discussion.

For example, a live ticker for a soccer portal app would be implemented. The question that arose out of the discussion was how often the ticker should be updated. The problem was discussed controversially, because the update rate should be quite high when a soccer match was in progress. On the other hand, if no soccer match was being broadcast then a low update rate would be sufficient. In the end, however, everyone agreed that if the refresh rate were implemented 'incorrectly', it would have a negative impact on the UX. If implemented 'correctly', it would impact the UX in a positive way.

The previous example shows that the same Epics,

depending on their implementations, can have both a positive and a negative impact on the UX. During Refinement, Epics tend to be evaluated and discussed based on their technical implementation. During the UX poker, the user is in the foreground and it is evaluated from his or her perspective.

C. Limitations

In this study, the use of UX Poker as a method was proposed and evaluated. Whether the estimated UX was actually achieved after development was not evaluated due to the time factor. This needs to be verified in further studies.

Furthermore, the study was only conducted in Germany. International studies should be conducted to exclude cultural and linguistic effects.

VI. Conclusions and Future Work

We have presented a method called 'UX Poker' for estimating the user experience for user stories or epics. The method aims to estimate the UX before implementing the user stories or epics. This has provided another way to sort or filter the Product Backlog in accordance with the estimation. We were able to evaluate this method in an initial study in workshops with 30 participants from four different companies.

The results showed that UX Poker could be used in a real-life application. All participants were able to use UX Poker on concrete examples. It was possible to estimate the possible UX to be achieved for all previously selected UX factors. The use of UX Poker also provided essential insights for the agile team as it took the user's perspective. As best practice, the use of personas in connection with UX Poker has proved to be useful. It helps the participants put themselves into the persona's shoes and assess the UX from the persona's perspective.

In our study, we conducted UX Poker as a separate Agile meeting. It would be necessary to evaluate whether this is appropriate or whether UX Poker can or should also be performed in the Planning Poker meeting. The combination of UX Poker and Planning Poker would save the Agile team another meeting.

In addition, it should be examined whether the UX maturity level of the team influences the results of UX Poker. The results of our study show that even participants with a developer background are pretty capable of successfully applying UX Poker. This suggests that the UX maturity level does not significantly influence the results. However, this needs to be confirmed in further studies.

Finally, it can be summarized that UX Poker is applicable in a real-life situation and that it helps to focus the agile team's attention on UX.

Acknowledgement

This research was partially funded by SocietySoft project (AT17_5904_USE) under Junta de Andalucia and Nico project (PID2019-105455GB-C31) of the Ministery of Science, Innovation and University, Spanish Government.

References

[1] J. Preece, Y. Rogers, H. Sharp, *Interaction design: Beyond human-computer interaction*. Chichester: Wiley, 4. ed. ed., 2015.

[2] ISO9241-210, "Ergonomics of human-system interaction - part 210: Human-centred design for interactive systems," 2020.

[3] ISO9241-220, "Ergonomics of human-system interaction - part 220: Processes for enabling, executing and assessing human - centred design within organizations," 2020.

[4] M. Hassenzahl, "The effect of perceived hedonic quality on product appealingness," *International Journal of Human-Computer Interaction*, vol. 2001, no. 13(4), pp. 481–499, 2001.

[5] N. Tractinsky, "Aesthetics and apparent usability," in *the SIGCHI conference*, 1997, pp. 115–122.

[6] M. Schrepp, J. Thomaschewski, "Design and validation of a framework for the creation of user experience questionnaires," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 7, pp. 88-95, 2019, doi: 10.9781/ijimai.2019.06.006.

[7] M. Rauschenberger, M. Schrepp, M. Perez- Cota, S. Olschner, J. Thomaschewski, "Efficient measurement of the user experience of interactive products. how to use the user experience questionnaire (ueq). example: Spanish language version," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 2, no. 1, p. 39, 2013, doi: 10.9781/ijimai.2013.215.

[8] M. Schrepp, A. Hinderks, J. Thomaschewski, "Design and evaluation of a short version of the user experience questionnaire (UEQ-S)," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 6, pp. 103-108, 2017, doi: 10.9781/ijimai.2017.09.001.

[9] J. Sauro, "Supr-q: A comprehensive measure of the quality of the website user experience," *Journal of Usability Studies*, vol. 2015, no. 10, pp. 68–86, 2015.

[10] M. Moshagen, M. T. Thielsch, "Facets of visual aesthetics," *International journal of human-computer studies*, vol. 68, no. 10, pp. 689–709, 2010, doi: 10.1016/j.ijhcs.2010.05.006.

[11] M. Schrepp, A. Hinderks, J. Thomaschewski, "Construction of a benchmark for the user experience questionnaire (ueq)," *International Journal of Interactive Multimedia and Artificial Inteligence*, vol. 4, no. 4, pp. 40–44, 2017, doi: 10.9781/ijimai.2017.445.

[12] A. Hinderks, M. Schrepp, F. J. Domínguez Mayo, M. J. Escalona, J. Thomaschewski, "Developing a ux kpi based on the user experience questionnaire," *Computer Standards & Interfaces*, vol. 65, pp. 38–44, 2019, doi: 10.1016/j.csi.2019.01.007.

[13] M. Schrepp, R. Otten, K. Blum, J. Thomaschewski, "What causes the dependency between perceived aesthetics and perceived usability?," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 6, pp. 78–85, 2020, doi: 10.9781/ijimai.2020.12.005.

[14] K. Schwaber, *Agile project management with Scrum*. Microsoft professional, Redmond, Wash.: Microsoft Press, 2004.

[15] D. J. Anderson, *Kanban: Successful evolutionary change for your technology business*. Sequim, Washington: Blue Hole Press, 2010.

[16] P. Serrador, J. K. Pinto, "Does agile work? — a quantitative analysis of agile project success," *International Journal of Project Management*, vol. 33, no. 5, pp. 1040–1051, 2015, doi: 10.1016/j.ijproman.2015.01.006.

[17] B. Boehm, R. Turner, "Using risk to balance agile and plan- driven methods," *Computer*, vol. 36, no. 6, pp. 57– 66, 2003, doi: 10.1109/MC.2003.1204376.

[18] K. Schwaber, J. Sutherland, *The Scrum Guide: The Definitive Guide to Scrum: The Rule of the Game*. 2020.

[19] A. Garcia, T. Silva da Silva, M. Selbach Silveira, "Artifacts for agile user-centered design: A systematic mapping," in *Proceedings of the 50th Hawaii International Conference on System Sciences (2017)*, Proceedings of the Annual Hawaii International Conference on System Sciences, 2017, Hawaii International Conference on System Sciences.

[20] M. Brhel, H. Meth, A. Maedche, K. Werder, "Exploring principles of user-centered agile software development: A literature review," *Information and Software Technology*, vol. 61, pp. 163–181, 2015, doi: 10.1016/j.infsof.2015.01.004.

[21] D. D. Brown, "Agile methods + ux = agile ux," in *Agile User Experience Design*, Elsevier, 2013, pp. 39–69, doi: 10.1016/B978-0-12-415953-2.00002-9.

[22] J. Choma, L. A. M. Zaina, D. Beraldo, "Userx story: Incorporating ux aspects into user stories elaboration," in *Human-Computer Interaction. Theory, Design, Development and Practice*, vol. 9731 of *Lecture Notes in Computer Science*, M. Kurosu Ed., Cham: Springer International Publishing, 2016, pp. 131–140, doi: 10.1007/978-3-319-39510-4_13.

[23] A. Joshi, N. L. Sarda, S. Tripathi, "Measuring effectiveness of hci integration in software development processes," *Journal of Systems and Software*, vol. 83, no. 11, pp. 2045–2058, 2010, doi: 10.1016/j.jss.2010.03.078.

[24] A. M. Moreno, A. Yagüe, "Agile user stories enriched with usability," in *Agile Processes in Software Engineering and Extreme Programming*, vol. 111 of *Lecture Notes in Business Information Processing*, Springer Berlin Heidelberg, 2012, pp. 168–176, doi: 10.1007/978-3-642- 30350-0_12.

[25] N. C. Haugen, "An empirical study of using planning poker for user story estimation," in *AGILE 2006 (AGILE'06)*, 2006, pp. 23–34, IEEE.

[26] M. Schrepp, J. Thomaschewski, "Handbook for the modular extension of the user experience questionnaire," 2019.

[27] A. Hinderks, F. J. Domínguez-Mayo, A.-L. Meiners, J. Thomaschewski, "Applying importance- performance analysis (ipa) to interpret the results of the user experience questionnaire (ueq)," *Journal of Web Engineering*, 2020, doi: 10.13052/jwe1540- 9589.1926.

[28] S. Gibbons, "Dot voting: A simple decision-making and prioritizing technique in ux," 2019. [Online]. Available: https://www.nngroup.com/articles/dot- voting/.

[29] D. Winter, C. Hausmann, E.-M. Schon, J. Thomaschewski, "Morphological analysis of the 'context of use' application of morphological analysis to the collaborative understanding of the context of use," in *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)*, 2020, pp. 1–6, IEEE.

[30] A. Cooper, R. Reimann, D. Cronin, *About face: Interface- und Interaction-Design*. Heidelberg and München and Landsberg and Frechen and Hamburg: mitp, 1 ed., 2010.

[31] D. Winter, E.-M. Schön, J. Thomaschewski, "Persona driven agile development: Build up a vision with personas, sketches and persona driven user stories," in *Information Systems and Technologies (CISTI)*, 2012.

Andreas Hinderks

Andreas Hinderks holds a PhD in Computer Science by University of Sevilla. He has worked in various management roles as a Business Analyst and a programmer from 2001 to 2016. His focus lay on developing user-friendly business software. Currently, he is a freelancing Product Owner, Business Analyst and Senior UX Architect. He is involved in research activities dealing with UX questionnaires, measuring user experience and User Experience Management since 2011.

Dominique Winter

Dominique Winter holds a Master of Science in Media Informatics from Emden/Leer University of Applied Sciences and a Master of Arts in Organisational Development from the TU Kaiserslautern. He works for various companies as a product development coach and supports them in improving their user orientation. He is also a doctoral student at the University of Siegen and conducts research on the topic of UX competence in and of organisations.

Francisco José Domínguez Mayo

Francisco José Domínguez Mayo received a PhD degree in computer science from the University of Seville, Seville, Spain, in July 2013. He is currently an associate professor with the Department of Computing Languages and Systems, University of Seville. He collaborates with public and private companies in software development quality and quality assurance. The focus of his interesting research is on the areas of continuous quality improvement and quality assurance on software products, and software development processes.

María José Escalona

María José Escalona received her PhD in Computer Science from the University of Seville, Spain in 2004. Currently, she is a Full Professor in the Department of Computer Languages and Systems at the University of Seville. She manages the web engineering and early testing research group. Her current research interests include the areas of requirement engineering, web system development, model-driven engineering, early testing and quality assurance. She also collaborates with public companies like the Andalusian Regional Ministry of Culture and Andalusian Health Service in quality assurance issues.

Jörg Thomaschewski

Jörg Thomaschewski received a PhD in physics from the University of Bremen (Germany) in 1996. He became a Full Professor at the University of Applied Sciences Emden/Leer (Germany) in September 2000. His teaching and research focus is on Human-Computer Interaction, UX-Management, Agile Software Development, and Requirements Engineering. Since 2012 he has been the head of the research group ‘Agile Software Development and User Experience’. Dr. Thomaschewski has extensive experience in user experience training, UX questionnaires, agile methods, IT analysis, and consulting.

Empirical Analysis of Ethical Principles Applied to Different AI Uses Cases

Alfonso José López Rivero[1], M. Encarnación Beato[1], César Muñoz Martínez[2], Pedro Gonzalo Cortiñas Vázquez[2] *

[1] Computer Science Faculty. Universidad Pontificia de Salamanca, Salamanca (Spain)
[2] Universidad Nacional de Educación a Distancia, Madrid (Spain)

Received 9 May 2022 | Accepted 7 July 2022 | Published 28 November 2022

Abstract

In this paper, we present an empirical study on the perception of the ethical challenges of artificial intelligence groups in the classification made by the European Union (EU). The study seeks to identify the ethical principles that cause the greatest concern among the population, analyzing these characteristics among different actors. The main study analyses the difference between Information and Communications Technology (ICT) professionals and the rest of the population. Along with this study, we conducted a gender study; in addition, we studied differences between university students, classified as future professionals who can work in Artificial Intelligence, and other university students. We believe that this work is a starting point for an informed debate in the scientific community and industry on the ethical implications of artificial intelligence based on the classification of ethical principles made by the EU, which can be extrapolated to any analysis carried out on the use of data to apply them in algorithms based on Artificial Intelligence.

Keywords

Artificial Intelligence, Ethic, Ethics Of AI, Digital Ethics, Trust, Digital Transformation.

DOI: 10.9781/ijimai.2022.11.006

I. Introduction

Artificial Intelligence (AI) is positioning itself as a strategic technology for the development of society in general and the economy in particular. Its use is increasingly widespread, covering all areas from healthcare, manufacturing, and security to agriculture and leisure. It is a life-changing technology with great prospects to benefit society as a whole; however, this growing use and potential for growth have also brought a recent concern to make AI systems trustworthy systems.

Work has been ongoing in recent years to make AI a trustworthy system: as early as 2018, the European Union (EU) published a coordinated plan on AI [1] in which one of the four basic pillars is the development of ethical and trustworthy AI. The achievement of trusted AI is closely linked to ethical AI that is based on fundamental values and rights such as human dignity and the protection of privacy. To this end, the EU's coordinated plan promoted the formation of an independent group of experts from different fields (academia, business, and society) to establish ethical guidelines for the development and use of AI systems. In 2019, this High-Level Expert Group on artificial intelligence (HLEG) published the document "Ethics guidelines for trustworthy AI" [2], with a vision centered on ethics as a fundamental pillar to guarantee trustworthy AI that benefits both human prosperity at the individual level and the common good of society.

* Corresponding author.
E-mail addresses: ajlopezri@upsa.es (A. J. López Rivero), ebeatogu@upsa.es (M. E. Beato), cmunoz@cee.uned.es (C. Muñoz Martínez), pcortinas@cee.uned.es (P. G. Cortiñas Vázquez).

This concern for the ethics of AI is not only present in Europe, in this same area there are other private initiatives such as the "Beneficial AI" conference organized by the Future of Life Institute in Asilomar (USA) in 2017, which was attended by the world's leading AI researchers and entrepreneurs and which resulted in the so-called "23 Asilomar principles" [3], or the work of the AI4People group [4] where a review of all the ethical principles proposed by different public and private initiatives is carried out. They compiled 47 principles and found that they can be grouped into the four basic principles used in bioethics, adding a fifth additional principle which would be explicability which includes responsibility. In another similar initiative by Jobin et al. [5], following the line of the previous work, a review of the main principles and guidelines for the ethics of AI was carried out, finding that there is currently a global convergence around five ethical principles (transparency, justice, and fairness, non-maleficence, responsibility, and privacy).

Another notable initiative is that of the United Nations Educational, Scientific and Cultural Organization (UNESCO), which, in 2020, set up an Ad Hoc Expert Group to draft recommendations on the ethics of artificial intelligence [6]. This document came out the same year intending to be a normative instrument based on international law that, with a primary focus on human dignity and human rights, could provide responsible guidance for AI technologies. Or the initiative similar to the one undertaken by the Organization for Economic Co-operation and Development (OECD) [7].

Other studies take a more philosophical view of AI, M. Graves [8] and J. F. Calderero [9] examine AI spiritually to suggest possible directions for ethics IA development. O. Krüger [10] goes further in his research by discussing the emergence of a superhuman computer intelligence for solving humanity's problems.

This is the context for this recently updated study, in which we conducted an empirical study with the aim of evaluating the public's perception of the ethical issues raised by the use of artificial intelligence, based on the ethical principles of reliable AI identified by the EU expert group. This work has chosen to work in a differentiated way with different population groups: firstly, with Information and Communications Technology (ICT) professionals, i.e. people who are directly involved in the conception, design, development, and implementation of AI-based systems, and therefore have a direct influence on the ethics of these systems. ICT professionals are a key group in the ethics of AI systems, so if these professionals have a lax perception of, or reflect little concern for, issues related to the ethics of AI, this could in principle negatively influence the ethics of AI systems they implement. A second, separate study group is university students in directly technology-related careers who will in the future be the designers and developers of AI systems. Finally, a third group is made up of people with no connection to the world of technology, users in general who use or will use products or services that make use of AI systems. The main objective is to assess the concern for AI ethics of these groups in a comparative way to see if there are significant differences between them that could influence the development of these systems in the future. In addition, differences by gender, differences between ICT and non-ICT university students, and age groups are analyzed. It is important to highlight the importance of the analysis by gender, as recent studies [11] show that only 22% of AI professionals are women, similar percentages are found among ICT students, so this gap will remain in the future; therefore, analyzing whether there are differences by gender is an issue that cannot be considered minor.

An analysis of all the recent bibliography highlights the study carried by T. Hagendorff and K. Meding [12], they analyze the 11,000 main papers from the main AI conferences worldwide, highlighting that the ethical implications are quite unknown, and although collaboration in publications between the academic and business world is increasing, it also highlights the difference with the articles presented by industry, where the ethical implications are analyzed less frequently than academic articles.

Another study [13] answered the question of how should ethics be implemented in autonomous systems and AI in our lives? It posits that the solution lies in philosophical conceptualization as a framework for forming a model for practical implementation of AI ethics. In the study, they conducted Systematic Mapping (SMS) on keywords used in AI and ethics to help identify, challenge, and compare the main concepts used in the current discourse of AI ethics. They analyzed over 1000 articles and discovered 37 keywords, considered the first step to guide and provide a direction for future research in the field of AI ethics.

Studies similar to these, in which a theoretical study on AI ethics is conducted, can be found in the literature, however, we have not found any empirical studies like the work presented in this article, which is why our study is novel and can be considered a starting point for an informed debate in the scientific community and industry. The ethical implications of artificial intelligence based on the classification of ethical principles made by the European Union can be extrapolated to any analysis carried out on the use of data to apply them in AI-based algorithms.

This article is organized as follows: Section II details the materials and methods used to carry out the study; section III contains the results obtained in the different studies; and finally, section IV sets out the conclusions obtained.

II. Methods

A. Materials

To carry out the study, we have designed a questionnaire of 15 questions, each of them specifically focused on one of the 5 ethical principles mentioned above. This questionnaire was validated by an interdisciplinary group of experts in the fields of ethics, technological ethics, and technology, from the Faculties of Philosophy and Theology of the Pontifical University of Salamanca. Concerning ethical considerations, the review was carried out by professors Dr. Marceliano Arranz Rodrigo and Dr. Gonzalo Tejerina Arias. To this main block of questions, another first block of questions was added to control the demographic characteristics of the subjects surveyed, including age category, level of studies, gender, and degree of knowledge of the technology on a scale of 1 to 10. Immediately after this, the block of questions on the main ethical challenges posed by artificial intelligence today was asked. These questions have been drawn up by experts in ethics in general, in technological ethics and technology, based on the four ethical principles identified by the HLEG group of experts [2], with the addition of a fifth, privacy, based on the grouping made by Jobin, Ienca, and Vayena [5]. These four principles are as follows: 1) Respect for human autonomy 2) Prevention of harm 3) Fairness 4) Explicability, 5) Privacy. Although all the questions are grouped here according to the ethical principle to which they belong, in the survey they were grouped to avoid reducing the bias caused by questions on the same issue in a row.

All the questions related to ethical aspects are written according to the same scheme: They describe a situation of use of artificial intelligence that poses an ethical dilemma and indicates their degree of acceptability on a scale from 1 (not at all acceptable) to 10 (completely acceptable), based on their own ethical decisions:

The situations related to each of these principles that make up the questionnaire are as follows:

1. *Respect for human autonomy*

   V1. The use of artificial intelligence to generate lifelike images and/or videos, and distribute them on social networks to create currents of opinion.

   V2. The use of artificial intelligence to serve as an electoral propaganda mechanism for parties on social networks.

   V3. The use of artificial intelligence that seeks to modify the consumption habits of the population.

   V4. The use of artificial intelligence in games that learns about the behaviour of players to increase the time spent playing the game.

   V5. The use of human-like robots to care for the elderly, capable of adapting to their needs, which could create affective dependence on the person being cared for.

2. *Prevention of harm*

   V6. Operate an autonomous driving vehicle that has not been sufficiently tested.

   V7. The use of artificial intelligence to integrate it into lethal autonomous weapons.

3. *Fairness*

   V8. The use of artificial intelligence where it is known that the data to be used for its learning is not of sufficient quality, with the risk that it learns badly.

   V9. The use of artificial intelligence for personnel selection, without human intervention and therefore objective, but the data could be biased in favour of men over women.

V10. The use of artificial intelligence to propose sentences in trials (in an objective way) in which the data from which the artificial intelligence learns could disadvantage certain races, social classes or groups.

4. *Explicability*

V11. The implementation of an artificial intelligence system in which control over the system does not depend on the human factor.

V12. The use of artificial intelligence to decide workers' salary supplements, knowing that it will not be possible to trace the reasons that lead the system to make such a decision.

5. *Privacy*

V13. The use of artificial intelligence through facial recognition to identify, record and learn about people's consumption habits, to stimulate the purchase of certain products.

V14. The use of artificial intelligence for video-surveillance of the public that, with the installation of cameras in the streets and facial recognition techniques, can obtain information on citizens by identifying their movements.

V15. The use of artificial intelligence to gather information on the tastes of the inhabitants of a house, using listening to virtual assistants and using it to make music recommendations.

B. Participants

The survey was applied to 457 people who fall into one of the following groups: ICT professionals, ICT students, non-ICT students, and general users. The sample of ICT and non-ICT students was obtained from university students in the academic years 2020/21 and 2021/22, while for ICT professionals the survey was distributed by the Professional Association of Computer Engineers of Castilla y León and AETICAL (Association of Technology Companies of Castilla y León). The users who completed the survey were selected by random sampling and the questionnaire was sent to them by email. Participation in the study was voluntary with informed consent; all data collected is considered confidential and only used for the study. The data have been deleted after the extraction of the aggregated data presented in this article.

For the surveys conducted, n=456, considering the worst-case scenario; p=q=0.5, where we have a non-finite population scenario where the total population is unknown and a confidence level of 95%, the sampling error is determined. The equation to determine the sample size can be found in many studies, in this research we have used the analysis of Louangrath, P.I., [14].

$$n = \left(\frac{Z*\sigma}{E}\right)^2 \quad (1)$$

Where Z, is the critical value for the confidence interval, in the study we have considered 95%, σ is the standard deviation, obtained by:

$$\sigma = \sqrt{p*q*n} \quad (2)$$

For the survey sample size of 456, the total error in the study is E=4.6%.

The socio-demographic characteristics of the sample are shown in Table I. The sample sizes are considered sufficiently large to conclude with a high degree of reliability.

TABLE I. Demographic Characteristics of the Sample

		n	%
Sexo	Male	288	63%
	Female	169	37%
Tipo	ICT professionals	94	20%
	ICT students	233	51%
	Non-ICT students	67	15%
	Others	63	14%
Edad	Under 30s	294	64%
	30-60 years old	102	22%
	Over 60 years old	61	14%

C. Tasks and Methods

1. Data Preprocessing

Before statistical analysis, data pre-processing needs to be performed to transform the raw data obtained into a "clean" and ordered dataset that allows for good statistical analysis.

Data pre-processing is collected from different sources [15], [16], a procedure for data pre-processing is as follows:

- Data cleaning: This stage deals with missing data, noise, outliers, and duplicate or incorrect records, and minimizes the introduction of bias into the database.
- Data integration: The extracted raw data may come from heterogeneous sources or be in separate datasets. This step reorganizes the different raw data sets into a single dataset containing all the information needed for the desired statistical analyses.
- Data transformation: this step translates and/or scales the initial variables into formats more useful for the statistical methods the researcher wishes to use.
- Data reduction: Once the dataset has been integrated and transformed, this step removes redundant records and variables and reorganizes the data in a more efficient and orderly manner for analysis.

During pre-processing, care must be taken not to accidentally introduce bias by modifying the dataset in a way that affects the outcome of statistical analyses, and to avoid arriving at statistically significant results by "trial and error" analysis on different versions of a pre-processed dataset.

In the study carried out, the following considerations have been taken into account concerning the points described above:

- Data cleaning: In the study that has been carried out, the ID has been checked to analyze duplicate data and, as for possible outliers of the variables analyzed, the first step was to check if there was a difference between the mean and median to detect possible values, in all cases the values were close; in addition, it was verified, by performing a scatter plot, if there were data out of range, in no case were outliers obtained.
- Data integration: In the analysis that has been carried out we have homogeneous sources of data as we have stratified data for the different groups that are analyzed: gender, workers/users, and ICT/non-ICT specialists.
- Data transformation: From the initial results obtained none of the variables followed a normal distribution, so the variables had to be transformed into normal distributions. This transformation is detailed in the following section.
- Data reduction: In the study presented, being a previously validated analysis, there are no redundant variables, and the data are ordered for the analysis.

2. Data Transformation. Normalisation of Variables

The quantitative data used for this study were obtained from a survey, collecting 15 quantitative variables that we wanted to measure concerning the perspective on ethics in artificial intelligence held by

different actors. In the first analysis of these variables, it was verified that they did not follow a normal distribution, as can be seen in Fig. 1 and Fig. 2, for the variable V1.

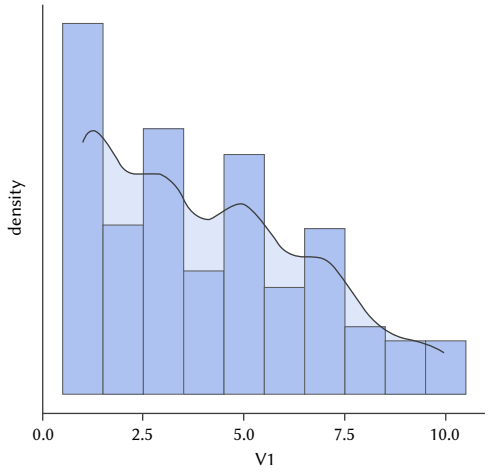


Fig. 1. Histogram for untransformed variable.

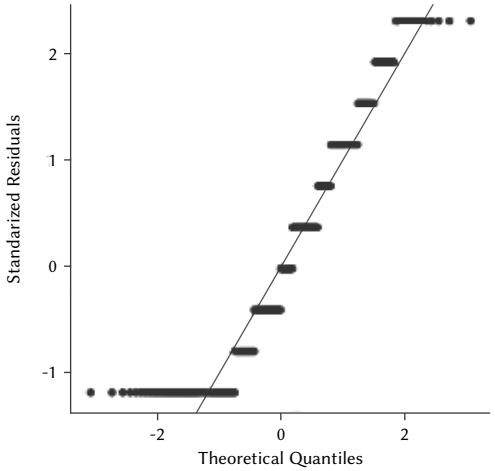


Fig. 2. Q-Q graphic diagram for untransformed variable.

In the histogram, it can be seen that the variable does not follow a normal distribution and if the Q-Q, Quantile-Quantile graph is performed, it is verified that the data do not follow a normal distribution function.

A similar process can be found in "Evaluating the normality of various quantitative data transformation procedures" by Dago et all [17], for which the following transformations are used: Z-score, Logarithmic, Adaptive Gamma Distribution, and Box-Cox.

3. Z-score Standardisation

In Z-transform data normalization, the x_i values of a variable V_i are normalized based on the mean and standard deviation of V [18]. Z-scores, or standard scores, indicate how many standard deviations and observations are above or below the mean. These scores are a useful way of putting data from different sources on the same scale. A value x_i of V is normalized to x_i' by calculating:

$$x_i' = \frac{x_i - \bar{x}}{\sigma_i} \tag{3}$$

Where $\bar{x}$ y σ_i are the mean and standard deviation of the variable V_i.

4. Logarithmic Transformation

The logarithmic transformation is very useful when results are influenced by independent factors, and in particular for transforming distributions with positive skewness, [19], the logarithmic transformation formula is given by:

$$x_i'' = \log x_i' \tag{4}$$

This logarithmic transformation is only valid for numbers greater than zero, to avoid having numbers less than zero when normalizing the variable, a constant is added to each value, in the study carried out the average value of the data has been added.

$$x_i'' = \log(x_i' + \bar{x}) \tag{5}$$

5. Adaptive Gamma Distribution

When the response variable is assumed to follow a parametric probability distribution, gamma in this study, the parameters of this distribution can be modeled, each independently, following linear, non-linear, non-parametric functions. This versatility makes the gamma function a suitable tool for modeling variables that follow a whole range of distributions: non-normal, asymmetric, or with non-constant variance [20].

6. Box-Cox Transformation

The Box-Cox transformation is a family of potential transformations of non-normal dependent variables, which allows transforming the variable to normal density functions, as this feature is a statistical assumption that allows using many more techniques for the analysis of the variables [21]. They take the idea of having a range of power transformations instead of the classical square root, logarithm, and inverse, available to improve the efficiency of normalization and variance equalization for variables with positive and negative skewness [22], [23]. The variable transformation has the form, where x are the original values we have in the sample, alpha is a transformation constant to avoid negative values when calculating the logarithm and lambda is the transformation parameter:

$$x_t^{\lambda} = \begin{pmatrix} \frac{(x_t)^{\lambda}-1}{\lambda} & if\ \lambda \neq 0 \\ \log(x_t + \alpha) & if\ \lambda = 0 \end{pmatrix} \tag{6}$$

In the study presented, the statistical analysis has been carried out with the R-based software package, JAMOVI [24], for the transformation of the variables the following equation has been followed, where K is a constant, the value 5 was taken as the initial scale is from 0 to 10, V_i is each of the 17 variables that have been analyzed, $\bar{V}$, σ_{v_i} are, respectively, the mean and the standard deviation of each variable.

$$V_i' = BOXCOX(GAMMA(LN\left(\frac{v_i - \bar{V}}{\sigma_{v_i}} + K\right)) \tag{7}$$

Fig. 3 and Fig. 4 show the result of the transformation performed for the first variable, V1, whose distribution function follows a normal density function. From the Q-Q plot it is found that the data fit the function.

If we consider the value of the Shapiro-Wilk statistic, we have that the transformed variable has a p-value of 0.762, which implies that the values fit a normal distribution.

III. Results

Firstly, we carried out a descriptive analysis of each of the variables related to ethical perception, grouped by ethical principle. This analysis is shown in Fig. 5, where each of the variables studied, grouped by ethical principle, is represented by a box plot.

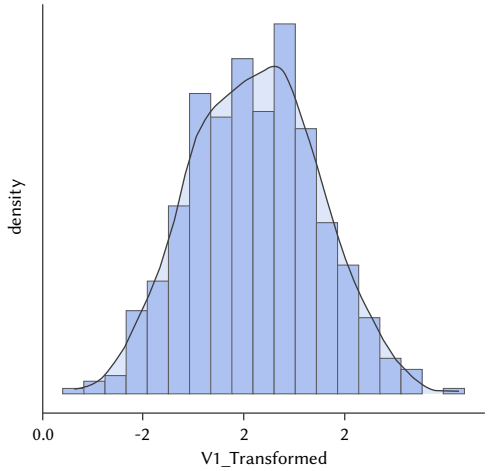


Fig. 3. Histogram for the transformed variable.

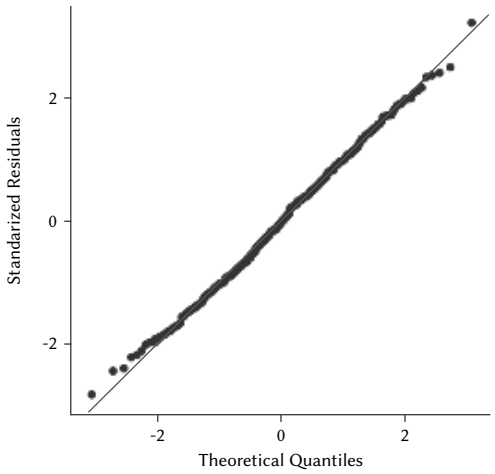


Fig. 4. Q-Q plot for the transformed variable.

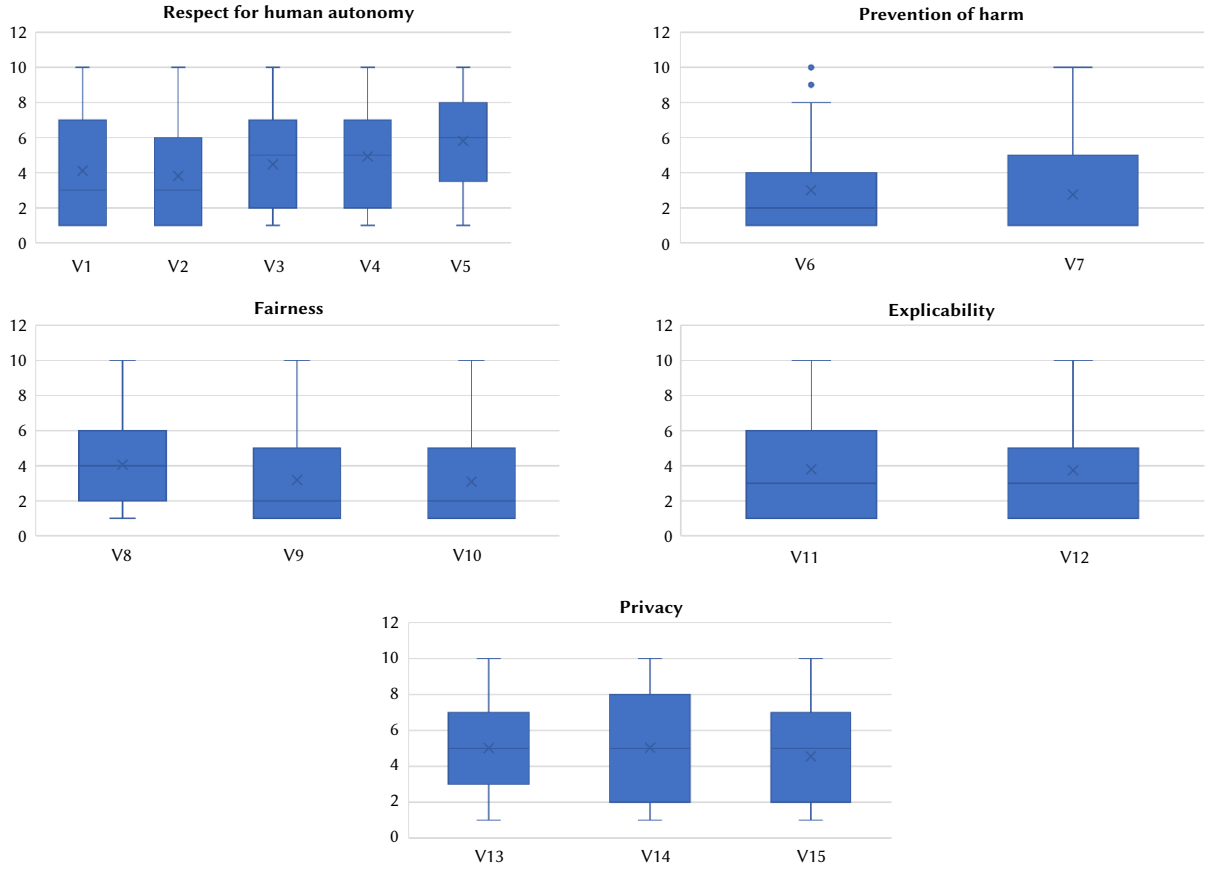


Fig. 5. Box plots showing the distribution of values for each variable and ethical principle.

Subsequently, an inferential analysis based on the p-value has been carried out. Before showing the results of these analyses, we will explain how to interpret the p-value and the effect size used.

The p-value measures the compatibility of a sample with a hypothesis, not that the hypothesis is true [25], as the p-value is a probability statement about the observed sample in the context of a hypothesis, not about the hypotheses being tested. This is why the American Statistical Association (ASA) [26], published an article in 2016 highlighting the misuse of p-values in statistical inference. Professor Cobb highlighted why we use p=0.05 as the cut-off value, citing numerous cases in which it is not used correctly. These errors in considering the interpretation of the p-value alone led Professor Nuzzo to publish an article on statistical errors in the scientific method [27].

Therefore, to test the significance of the hypotheses, in addition to the p-value, other parameters must be considered, and in this analysis, we have considered the effect size, a concept introduced by Cohen [28].

Effect sizes, Cohen's d, are the most important result of empirical studies [29], used to describe the standardized mean difference of an effect. This value can be used to compare effects across studies, even when the dependent variables are measured in different ways. A commonly used interpretation is to refer to effect sizes as small (d = 0.2), medium (d = 0.5) and large (d = 0.8) according to the benchmarks suggested by Cohen [28].

TABLE II. Averages. Users (1) Vs. ICT Professionals (2)

		Mean		Median		Standard deviation	
		1	2	1	2	1	2
Respect for human autonomy	V1	4.58	2.29	5	1	2.87	2.19
	V2	4.17	2.39	4	1	2.76	2.26
	V3	4.84	3.09	5	2	2.69	2.50
	V4	5.14	4.02	5	3.5	2.95	2.80
	V5	5.74	6.06	6	7	2.77	2.73
Prevention of harm	V6	3.09	2.69	2	2	2.59	2.41
	V7	2.95	2.04	1	1	2.69	2.05
Fairness	V8	4.25	3.28	4	3	2.57	2.50
	V9	3.29	2.82	2	1	2.68	2.69
	V10	3.21	2.57	2	1	2.70	2.49
Explicability	V11	3.86	3.54	3	3	2.70	2.60
	V12	3.83	3.38	3	3	2.49	2.66
Privacy	V13	5.35	3.71	5	3	2.64	2.71
	V14	5.30	4.00	5	3	2.87	2.84
	V15	4.86	3.40	5	2.5	2.89	2.65

TABLE III. Users in General Vs. ICT Professionals

				95% Confidence Interval	
		p	Effect Size	lower	upper
Respect for human autonomy	V1	**0.025**	0.2603	0.03250	0.4877
	V2	**0.023**	0.2631	0.03532	0.4906
	V3	**0.001**	0.3827	0.15392	0.6110
	V4	0.561	0.0673	-0.15962	0.2941
	V5	**0.037**	-0.2423	-0.46960	-0.0146
Prevention of harm	V6	0.645	-0.0534	-0.28022	0.1735
	V7	0.502	0.0778	-0.14919	0.3046
Fairness	V8	0.769	0.0340	-0.19287	0.2608
	V9	0.445	0.0884	-0.13858	0.3153
	V10	0.384	0.1009	-0.12614	0.3277
Explicability	V11	0.740	0.0385	-0.18841	0.2653
	V12	0.412	0.0951	-0.13188	0.3220
Privacy	V13	**0.042**	0.2360	0.00838	0.4633
	V14	0.372	0.1035	-0.12352	0.3304
	V15	**0.016**	0.2802	0.05231	0.5078

TABLE IV. Average. Men (1) Vs Women (2)

		Mean		Median		Standard deviation	
		1	2	1	2	1	2
Respect for human autonomy	V1	4.09	4.14	3.00	4	2.98	2.76
	V2	3.84	3.75	3.00	3	2.77	2.76
	V3	4.42	4.59	5.00	5	2.76	2.72
	V4	5.52	3.87	6.00	4	2.93	2.69
	V5	6.29	4.99	7.00	5	2.58	2.87
Prevention of harm	V6	3.24	2.62	2.00	1	2.65	2.35
	V7	3.07	2.25	1.00	1	2.77	2.17
Fairness	V8	4.23	3.75	4.00	3	2.51	2.68
	V9	3.58	2.52	2.50	1	2.80	2.34
	V10	3.39	2.56	2.00	1	2.81	2.33
Explicability	V11	4.21	3.08	4.00	2	2.68	2.54
	V12	4.04	3.22	4.00	2	2.55	2.42
Privacy	V13	5.28	4.57	5.00	4	2.80	2.58
	V14	5.10	4.91	5.00	5	3.00	2.75
	V15	4.67	4.37	5.00	4	2.97	2.76

TABLE V. Difference by Gender Males Vs. Females

				95% Confidence Interval	
		p	Effect Size	lower	upper
Respect for human autonomy	V1	0.429	-0.0767	-0.267	0.1134
	V2	0.857	-0.0175	-0.207	0.1724
	V3	**0.036**	-0.20401	-0.3945	-0.0132
	V4	0.559	0.05667	-0.1334	0.2466
	V5	**<.001**	0.36639	0.1738	0.5583
Prevention of harm	V6	0.771	0.02820	-0.1618	0.2181
	V7	**0.014**	0.24028	0.0491	0.4310
Fairness	V8	0.132	0.14613	-0.0443	0.3363
	V9	0.085	0.16746	-0.0231	0.3577
	V10	0.654	0.04342	-0.1466	0.2333
Explicability	V11	0.267	0.10769	-0.0825	0.2977
	V12	0.371	0.08679	-0.1033	0.2768
Privacy	V13	0.178	0.13063	-0.0597	0.3207
	V14	0.983	-0.00209	-0.1920	0.1878
	V15	0.690	0.03872	-0.1513	0.2286

Concerning the analyses carried out, Table II shows the average scores for each question for ICT professionals, people who are working in this sector, compared to the rest of the respondents (General users: 1, professionals: 2).

With these data, we have analyzed the differences in ethical perceptions in each of the groups. The result of these differences for each of the ethical principles is shown in Table III. In this table, and all the equivalent tables in this section, we have marked in black the questions where we found significant differences between the groups.

In the second analysis, we have carried out a gender study to see if there are differences between men and women concerning the perception of ethical principles. The results of the averages obtained and the statistical analysis of differences are shown in Table IV and Table V, respectively.

Next, we have carried out an analysis to determine whether differences are found between students in careers directly related to technology versus non-technology students.

Bearing in mind that technology students will be future ICT professionals, the aim is to analyze whether their perception of the ethics of AI is different from that of other university students in non-technology-related subjects. The results of the analysis are shown in Table VI and Table VII.

Finally, we have carried out an analysis by age, and we have considered whether there are differences between young people

TABLE VI. AVERAGE NON-ICT UNIVERSITY STUDENTS (1) VS. ICT STUDENTS (2)

		Mean		Median		Standard deviation	
		1	2	1	2	1	2
Respect for human autonomy	V1	5.03	4.85	5	5	2.65	2.95
	V2	4.42	4.54	4	5	2.55	2.80
	V3	5.30	4.94	5	5	2.55	2.71
	V4	4.90	5.88	5	6	2.74	2.82
	V5	4.52	6.37	4	7	2.73	2.61
Prevention of harm	V6	2.73	3.45	1	2	2.42	2.74
	V7	2.15	3.54	1	2	2.07	2.91
Fairness	V8	4.48	4.51	5	4	2.59	2.52
	V9	2.55	3.87	2	3	2.03	2.88
	V10	3.22	3.57	2	2	2.71	2.81
Explicability	V11	3.52	4.37	3	4	2.39	2.75
	V12	3.46	4.33	3	4	2.34	2.51
Privacy	V13	5.54	5.80	5	6	2.40	2.59
	V14	5.87	5.43	6	6	2.91	2.93
	V15	5.12	5.06	5	5	2.83	2.96

TABLE VII. NON-ICT VS. ICT UNIVERSITY STUDENTS

				95% Confidence Interval	
		p	Effect Size	lower	upper
Respect for human autonomy	V1	0.609	-0.07102	-0.3427	0.20122
	V2	0.300	-0.14399	-0.4163	0.12935
	V3	0.839	0.02817	-0.2437	0.29981
	V4	**0.048**	-0.27480	-0.5495	0.00191
	V5	**0.003**	-0.41330	-0.6925	-0.13117
Prevention of harm	V6	0.944	-0.00975	-0.2814	0.26199
	V7	0.889	0.01933	-0.2525	0.29097
Fairness	V8	0.551	-0.08271	-0.3545	0.18967
	V9	0.072	-0.25059	-0.5247	0.02537
	V10	0.455	-0.10377	-0.3757	0.16890
Explicability	V11	0.716	0.05041	-0.2216	0.32206
	V12	**0.020**	0.32495	0.0465	0.60105
Privacy	V13	0.555	-0.08186	-0.3536	0.19051
	V14	0.725	0.04886	-0.2232	0.32051
	V15	0.645	0.06389	-0.2083	0.33557

TABLE VIII. AVERAGES OVER 60 YEARS OLD (1) VS. UNDER 30 (2)

		Mean		Median		Standard deviation	
		1	2	1	2	1	2
Respect for human autonomy	V1	2.33	4.36	1	4.00	2.34	2.93
	V2	2.43	3.97	1	3.00	2.47	2.75
	V3	2.97	4.70	2	5.00	2.52	2.66
	V4	3.98	4.93	3	5.00	2.95	2.92
	V5	5.97	5.60	7	6.00	2.74	2.77
Prevention of harm	V6	2.62	2.93	1	2.00	2.36	2.49
	V7	2.08	2.63	1	1.00	2.25	2.48
Fairness	V8	3.20	4.09	3	4.00	2.55	2.61
	V9	2.95	2.95	1	2.00	2.85	2.48
	V10	2.59	3.00	1	2.00	2.64	2.61
Explicability	V11	3.26	3.73	3	3.00	2.48	2.66
	V12	3.28	3.68	2	3.00	2.65	2.51
Privacy	V13	3.69	5.11	3	5.00	2.79	2.66
	V14	3.77	5.17	3	5.00	2.72	2.84
	V15	3.13	4.79	2	5.00	2.72	2.85

TABLE IX. OLDER VS. YOUNGER

				95% Confidence Interval	
		p	Effect Size	lower	upper
Respect for human autonomy	V1	0.340	-0.1344	-0.4106	0.1430
	V2	0.090	-0.2390	-0.5170	0.0410
	V3	**<.001**	-0.5285	-0.8179	-0.2350
	V4	0.882	0.0209	-0.2549	0.2966
	V5	0.055	0.2707	-0.0103	0.5496
Prevention of harm	V6	0.183	0.1877	-0.0909	0.4647
	V7	0.604	0.0730	-0.2034	0.3488
Fairness	V8	0.327	0.1380	-0.1394	0.4143
	V9	0.858	0.0252	-0.2507	0.3009
	V10	0.251	-0.1616	-0.4382	0.1163
Explicability	V11	0.787	-0.0381	-0.3137	0.2379
	V12	0.919	-0.0143	-0.2900	0.2615
Privacy	V13	0.112	-0.2241	-0.5018	0.0555
	V14	0.168	-0.1944	-0.4715	0.0843
	V15	**0.013**	-0.3520	-0.6335	-0.0678

(under 30 years old) and older people (over 60 years old). We have carried out the analysis with this casuistry to eliminate the bias that could be introduced by ICT professionals, all of whom are middle-aged so that professionals are excluded from this analysis by age. The results of this study are shown in Table VIII and Table IX.

IV. DISCUSSION

We will now analyze the results obtained. First, we will focus on the descriptive analysis of the different situations on technological ethics. As indicated, the situations presented all pose some ethical challenge related to AI and the respondent is asked to rate them from 1 (not at all acceptable) to 10 (totally acceptable) using their own ethical decisions. The numerical scale of 1 to 10 has been chosen because of its similarity to the ratings, so we can understand that values lower than 5 indicate non-acceptance of the situation presented and higher than 5 indicate acceptance, although the numerical value will give us an idea of the magnitude of acceptance.

If we look at Fig. 1 and analyze the data represented there descriptively, we can see that in the analysis of averages there are only three of the 15 situations presented that obtain an average score higher than 5, these being variables V5, V13, and V14, with V5 standing out in particular with an average of 5.81. Variable V5, which falls within the ethical principle of respect for human autonomy, is related to the use of robots with a human appearance for the care of the elderly. The use of these robots for the care of the elderly means that it is

valued positively and very highly with respect to the rest, despite the fact that, as reflected in the situation presented, it may create affective dependence in a particularly sensitive group such as the elderly. On the other hand, variables V13 and V14, related to the ethical principle of privacy, reflect situations in which facial recognition techniques are used, on the one hand, to find out consumer habits and stimulate purchases (V13) and, on the other, to identify citizens' movements for street surveillance (V14). In principle, this reflects the fact that citizens are not as concerned about their privacy as they are about the data protection and privacy regulations that exist in all countries.

On the other side of the spectrum, the variables V7 (2.77), V6 (3.01), V10 (3.08), and V9 (3.19), all with values close to or below 3, stand out for their low average scores. The scores show concern for the safety/reliability of these systems when they may affect human life, as they are situations of AI use in autonomous weapons and self-driving vehicles, although it should be noted that there may have been a bias due to the rejection of anything related to weapons. On the other hand, situations V10 and V9 are found within the principle of fairness and are closely linked to the rights to non-discrimination, solidarity, and justice, the second principle that causes the most misgivings. It reflects situations in which AI is used for decision-making in the knowledge that bias in the data may disadvantage women against men or certain races (if used to determine convictions in trials). However, when the situation raised about fairness is more general, i.e., in V5, where it is only mentioned that the data used is not of sufficient quality without indicating what problems this may entail, the average score of the respondents is significantly higher than 4.05, perhaps because many of the respondents do not manage to assess the problems that the lack of quality may bring.

In the following, we will analyze the results obtained from the comparative analyses. The main idea behind this study is to assess whether there are differences between the ethical perception of people who are working in technology (professionals) and the rest. We believe that the ethical perception of professionals can influence the ethics of the AI products they develop and, therefore, it is important that their concern for ethical issues is high. If we analyze the results obtained in Table III, we observe that all the differences found between the two groups are within the ethical principle of respect for autonomy, strongly associated with the right to human dignity and liberty, specifically in variables V1, V2, V3, and V5. In the first three, situations are presented that make use of AI for different types of manipulation. Thus, in V1, AI is used to create real-looking images or videos that can create currents of opinion, in V2, AI is used by political parties for their propaganda, and in V3, AI is used to modify the population's consumption habits. The effect sizes (Cohen's d) in them (0.26, 0.26, and 0.38) indicate a small but significant effect size. Furthermore, if we analyze the sign and mean scores of these two groups on each variable (Table II), we can observe that the mean scores on these variables are much lower in the case of professionals. In other words, professionals are less permissive on issues related to human autonomy. Except in variable V5 which has to do with the use of IA for the care of the elderly, where the score is reversed, with professionals being more permissive. In this variable, the effect size is -0.24.

If the descriptive data collected in Table II are analyzed in general, it can be seen that for all the variables analyzed except for the care of the elderly, the professionals are on average less permissive than the rest, which would indicate that their concern for the ethics of AI is greater.

We have also carried out a comparative analysis by gender, men, and women, in order to analyze whether there are differences between them in terms of concern for the ethics of AI. The results of the analysis are reflected in Table V. If we look at the table, we find differences in variables V3 and V5 related to human autonomy and variable V7 related to prevention of harm. However, although the order of magnitude of the effect size is similar, the sign is not the same. Thus, in V3 where AI is used to modify the consumption habits of the population, women are more permissive than men (Cohen's d -0.20), while in the other two variables where there are differences, V5 (use for elderly care) and V7 (to integrate it into autonomous weapons), women are less permissive than men, reflecting in both scores lower than those of men. Thus, the average score (Table IV) for women in the question on the use of AI in weapons (V7) is 2.25 compared to 3.07 for men, being in both cases far from approval, which would be 5. And in the question on the use of AI for the care of the elderly that can cause emotional dependence (V5), a lower score is also obtained in the case of women, the average for them is 4.99 compared to 6.29 for men. In any case, this is the question that practically passes in both cases.

Once the gender analysis and the analysis of professionals versus the rest were done, an analysis was carried out to look for differences between university students in ICT-related degrees and other university students in other fields. The idea of this analysis was to see whether ICT students were, already in their university years, more aware of the ethical issues involved in the use of AI. The results of this analysis are reflected in TABLE VI and TABLE VII. If we look at TABLE VI, we see that there are significant differences in only three variables (V4, V5, and V15), with the sign the negative effect in all of them, which indicates that ICT students value these situations with higher scores, i.e., they are more permissive or have fewer ethical problems with these situations. The results for these three variables do not ratify the initial hypothesis that ICT students were more aware of issues related to the ethics of AI. Specifically, the situations raised refer to the use of AI in games to increase playing time (V4 human autonomy), that of caring for the elderly, and finally V15 in which the use of AI by virtual assistants to make recommendations about music (privacy) is raised. Analyzing these results and looking at the percentage of men and women in both categories, we find that in the case of ICT students the percentage of women is 25% while among non-ICT students it is 63%. To see if this fact could have an influence, we carried out a gender analysis only with the students, only finding differences in favor of women (less permissive) in question V5, the one related to caring for the elderly. Therefore, we could indicate that only variables V4 and V15 have a different behavior between ICT and non-ICT students, with NON-ICT students being less permissive in the use of AI in the aforementioned situations.

Finally, we have carried out an analysis by age, in this case, we have carried out an analysis of young people (under 30) versus those over 60, in order to see if age is a factor that can influence the perception of ethics. Middle-aged subjects have been excluded from this analysis in order to eliminate the effect, already studied at the beginning of this section, of ICT professionals, as they all belong to this intermediate age group. The results of this analysis are reflected in TABLE VIII and TABLE IX, in which differences are only found in variables V3 (seeks to modify consumption habits) and V15 (use of AI by virtual assistants for music recommendations), with the assessment of the over-60s being more restrictive in both variables, with an average of 2.97 compared to 4.70 for young people for variable V3 and 3.13 compared to 4.79 for variable V15. However, we did not find any difference in the rest of the variables studied, in which age has no influence, so it is not considered a determining factor.

V. Conclusion

Following the field study in which we evaluated the population's perception of the ethics of AI systems, the first conclusion we draw from this study is that, in general, the population does not consider the use of AI to be acceptable in situations in which an ethical problem arises and, therefore, we can affirm that there is a high level of

concern among the population about the ethics of these systems. This affirmation is corroborated by the average obtained in the responses, taking into account all the variables, which is 3.79, well below the pass mark. It should be remembered that in the situations presented, an ethical problem related to one of the ethical principles identified by the EU's high-level expert group on artificial intelligence [2] is presented and the degree of acceptability is indicated, on a scale from 1 Not at all acceptable to 10 Completely acceptable.

On the other hand, if we go into the concern of each of the five ethical principles, we find that the ones that raise the most concern are those of prevention of harm and fairness, the first of which is strongly linked to the protection of physical integrity, so the result is to be expected. The second, fairness, is related to ensuring that individuals and groups are free from unfair bias, discrimination, and stigmatization. The study also highlights the good reception of AI when it is used for the care of the elderly, even though it may create emotional dependence, with an average score of 5.81, as well as the average scores above 5 in the situations raised concerning video surveillance, in which people's privacy may be violated.

As far as ICT professionals are concerned, their concern about the ethics of AI is found to be higher than the rest of the population, indicating that they are even less permissive than the general population. This is a promising scenario as they are the first guarantors of the ethics of AI systems and project good prospects for the future in this area.

As for ICT university students, we found hardly any differences concerning other university students, which indicates that, although professionals in these fields are more aware of the ethics of AI, this is not the case when they are in their university years, where we found no differences.

Finally, we would like to point out that this is a field study carried out in Spain, so future lines of work will focus on extending it to other countries, as well as carrying out longitudinal work that will allow us to monitor and analyze the evolution of concern for the ethical issues of AI as the technology evolves and becomes more present in society.

Acknowledgment

We would like to thank the experts in ethics at the Pontifical University of Salamanca, Dr. Marceliano Arranz Rodrigo, retired professor of the Faculty of Philosophy and Dr. Gonzalo Tejerina Arias, professor of the Faculty of Theology, for their contributions and comments on the design of the survey for data collection.

References

[1] European Commission. Artificial Intelligence for Europe, 2018. Accessed: Jun. 24, 2022. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN.

[2] European Commission, Directorate-General for Communications Networks, Content and Technology. Ethics guidelines for trustworthy AI, 2019, Accessed: Jun. 24, 2022. [Online]. Available: https://data.europa.eu/doi/10.2759/346720.

[3] Future of Life Institute. ASILOMAR principles. Principles developed in conjunction with the 2017 Asilomar conference, 2017. Accessed: Jun. 24, 2022. [Online]. Available: https://futureoflife.org/2017/08/11/ai-principles/

[4] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U, Pagallo, F. Rossi, B, Schafer, P. Valcke, and E. Vayena, "AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations," *Minds and machines*, vol. 28 no. 4, pp. 689-707, 2018.

[5] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, vol. 1 no. 9, pp. 389-399, 2019.

[6] UNESCO Ad Hoc Expert Group (AHEG). First Draft of The Recommendation on the Ethics of Artificial Intelligence, 2020. Accessed: Jun. 24, 2022. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000373434

[7] OCDE. Recommendation of the Council on Artificial Intelligence. 2019. Accessed: Jun. 24, 2022. [Online] https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449?_ga=2.248246209.790951165.1621927053-1872220198.1621927053

[8] M. Graves, "Emergent Models for Moral AI Spirituality," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7 no. 1, pp. 7-15, 2021.

[9] J. F. Calderero, "Artificial Intelligence and Spirituality," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7 no. 1, pp. 34-43, 2021.

[10] O. Krüeger, "The Singularity is near!. Visions of Artificial Intelligence in Posthumanism and Transhumanism," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7 no. 1, pp. 16-23, 2021.

[11] S. Duke, "Will AI make the gender gap in the workplace harder to close?" World Economic Forum. 2018. Accessed: Jun. 24, 2022. [Online] https://www.weforum.org/agenda/2018/12/artificial-intelligence-ai-gender-gap-workplace/

[12] T. Hagendorff and K. Meding, "Ethical considerations and statistical analysis of industry involvement in machine learning research," *AI & SOCIETY*, pp. 1-11, 2021.

[13] V. Vakkuri and P. Abrahamsson, "The Key Concepts of Ethics of Artificial Intelligence," in *IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*, 2018, pp. 1–6.

[14] P. Louangrath, "Sample Size Calculation for Continuous and Discrete Data," *International Journal of Research & Methodology in Social Science*, vol. 5 no. 4, pp. 44-56, 2019.

[15] B. Malley, D. Ramazzotti and J. Wu, "Data Pre-processing,". *Secondary Analysis of Electronic Health Records*, pp. 115-141, 2016.

[16] X. Chu, I. F. Ilyas, S. Krishann and J. Wang, "Data Cleaning: Overview and Emerging Challenges," in *Proceedings of the 2016 International Conference on Management of Data, (SIGMOD)*, 2016, pp. 2201–2206.

[17] D. N. Dago, G. A. J. Kablan, K. A. Alui, H. D. Lallié, D. Dagnogo, N. Diarrassouba and M. Giovanni, "Normality Assessment of Several Quantitative. Data Transformation Procedures," *Biostatistics and Biometrics*, vol. 10 no. 3, pp. 51-65, 2021.

[18] D L. A. Shalabi, Z. Shaaban and B. Kasasbeh, "Data Mining: A Preprocessing Engine," *Journal of Computer Science*, vol. 2 no. 9, pp. 735-736, 2006.

[19] R. M. West, "Best practice in statistics: The use of log transformation," *Annals of Clinical Biochemistry*, vol. 59 no. 3, pp. 162-165, 2022.

[20] A. A. Al-Babtain, I. Elbatal, C. Chesneau and F. Jamal, "Box-Cox Gamma-G Family of Distributions: Theory and Applications," *Mathematics,* vol. 8 no. 10, pp. 1801, 2020.

[21] G. Box and D. R. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society: Series B (Methodological),* vol. 26 no. 2, pp. 211-243, 1964.

[22] J. Osborne, "Improving your data transformations: Applying the Box-Cox transformation," *Practical Assessment, Research, and Evaluation,* vol. 15 no. 1, pp. 12, 2010.

[23] N. Ishak and S. Ahmad, "Estimating Optimal Parameter of Box-Cox Transformation in Multiple Regression with Non-normal Data," in *Conference on Science, Technology and Social Sciences (RCSTSS 2016)*, Singapore, 2018, pp. 1039–1046.

[24] The jamovi project (2021). *jamovi* (Version 2.2) [Computer Software]. Retrieved from https://www.jamovi.org

[25] N. Altman and M. Krzywinski, "Points of significance: interpreting P values," *Nature Methods,* vol. 14 no. 3, pp. 213-215, 2017.

[26] R. L. Wasserstein and N. A. Lazar, "The ASA statement on p-values: context, process, and purpose," *The American Statistician,* vol. 70 no. 2, pp. 129-133, 2016.

[27] R. Nuzzo, "Statistical errors," *Nature,* vol. 506 no. 7487, pp. 150-152, 2014.

[28] J. Cohen, *Statistical power analysis for the behavioral sciences*, Lawrence Erlbaum Associates, Publishers, 1988.

[29] D. Lakens, "Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs," *Frontiers in psychology,* vol. 4, pp. 863, 2013.

Alfonso José López Rivero

Alfonso José López Rivero. PhD in Computer Science. He is a professor, since 1996, and member of the research group GESTICON (Ethical and Technological Management of Knowledge) at the Pontifical University of Salamanca, UPSA, (Spain). He is a member of the organizing and scientific committee of several international symposia and co-author of articles published in several recognized journals, workshops and symposia. He has been Dean of the Faculty of Computer Science and Director of the Office for the Transfer of Research Results at UPSA.

M. Encarnación Beato

Mª Encarnación Beato (PhD.). Received a PhD. in Computer Science from the University of Valladolid in 2004. She is professor at the Universidad Pontificia de Salamanca (Spain) since 1997. At present she is a member of the MARATON (Mobile Applications, inteRnet of things, dAta processing, semanTic technologies, OpeN data) research group at the Universidad Pontificia de Salamanca. She has been a member of the organizing and scientific committee of several international symposiums and has co-authored papers published in a number of recognized journals, workshops and symposiums.

César Muñoz

César Muñoz. PhD in European Union. He is an associate professor at the National Distance Education University (Spain). His research area focuses on transport and tourism economics. He's especially interested in exploring the application of ICT in these sectors. He has published several articles in scientific journals, including, International Journal of Tourism Research or Tourism Geographies.

Pedro Gonzalo Cortiñas Vázquez

Pedro Cortiñas Vázquez holds a PhD in Economics and Business Administration and is Professor of Applied Economics and Statistics at the Universidad Nacional de Educación a Distancia. Master in Commercial Management and Marketing by ESIC, University Expert in Multivariate Analysis and in Market Research and Opinion Polls, Specialist in Macroeconomics by The MIT Sloan School of Management.

Marketing Intelligence: Boom or Bust of Service Marketing?

Jan Lies*

FOM University of Applied Science, Dortmund (Germany)

Received 28 January 2021 | Accepted 11 April 2022 | Published 3 October 2022

Abstract

Marketing intelligence fosters two major developments within digital service marketing. On the one hand, a boom of services seems to have evolved, accelerated by the opportunities of marketing intelligence. It has contributed to the optimization of customer experiences, e.g., supported by mobile, personalized, and customized marketing services. On the other hand, (digital) self-services are likely to pervert the term "service". Lifecycle marketing, including annoying marketing communication in real-time, automated price adjustment and programmatic advertising based on artificial intelligence, affects the vision of fully standardized marketing automation. Additionally, there are incentives to pollute the digital information in order to manufacture opinions. Fake news is one popular example. This leads to the (open) question if marketing intelligence means service boom or bust of marketing. This contribution aims to elaborate the boom-and-bust aspects of marketing intelligence and suggests a trade-off. The method applied in this paper will be a descriptive and conceptual literature review, through which the paradigmatic thoughts will be juxtaposed from the perspective of service.

Keywords

Digitality, Marketing Intelligence, Servitization.

DOI: 10.9781/ijimai.2022.10.001

I. Introduction

MARKETING evolves. Contemporarily, a marketing 5.0 debate is underway, integrating the "values view" of marketing 3.0 and a "digitization view" of marketing 4.0. In retrospect, this debate ideally shows a development that focuses on human-centric progress through digitalization [1]:

- Marketing 1.0: The core competence of marketing is focused on a product and its distribution (since the 1950´s).
- Marketing 2.0: Marketing shifts to the consumer (beginning in the 1970´s).
- Marketing 3.0: Human centered marketing with the focus on people (from the 1980´s onwards).
- Marketing 4.0: The focus here is on digitalization (from approx. 2010 onwards)
- Marketing 5.0 is characterized by the integration of human centricity and digitization (forthcoming).

Marketing 1.0 originally and traditionally is product related. The core "p" of the traditional marketing mix means "product". Thus, a "breaking free of product marketing" and a shift to values became necessary [2]. The notion of servitization of marketing and beyond evolves [3]. It requires the opening up of the traditional "4p" (product, price, place, promotion) with additional "p"´s, specifically "persons" (e.g. sellers, vendors, stakeholders), their "problems" (e.g. needs, values) and processes (e.g. interactions, dialogues) in regards to services [4]. Meanwhile, the interdependence of service quality and customer satisfaction are discussed as a core concept of the marketing management [5]. Marketing intelligence contributes to marketing 4.0 and shapes digitized data-driven marketing. The term "digital marketing" has evolved with the emergence of the instruments of digital communication, channels and processes [6]. It supports marketing decisions by the analysis and application of digital customer data – even in real time. Moreover, web-based cloud branding, search engines, recommender and customer review systems, as well as the semantic web are part of the digitized marketing evolution. They enable an "explosion in the number of technology solutions" [7], i.e., marketing services.

To characterize services today, two dominating logics have to be mentioned: the goods-dominant logic views services in terms of an intangible good. The service-dominant logic views services as the fundamental objective of economic exchange which needs to impact marketing thinking [8]. Interactive relationships shape this view of service thinking. The value-in-use notion not only includes information, but also product transfer, upgrades, problem solving, etc. [9]. This service-dominant logic of corporations is not reduced to marketing but is "an all pervasive part of their strategic mission and corporate planning" [10]. This customer-driven development is called servitization. Here, this term is understood as a portmanteau of "service" and "digitization" that ideally impact three dimensions of services: 1. Services as added values of market outcomes; 2. Intangible outcomes; 3. Services as logics dominating corporate behaviour.

Thus, the idea of services means more than the shift from producing (tangible) goods to (intangible) added services. A constituting characteristic of services is the participation in the production. It has become an impacting factor of the dominant logic of service marketing. Thus, service marketing fosters mutually beneficial relationships between a firm and its customer(s) and, if possible, also society [11].

* Corresponding author.
E-mail address: jan.lies@jan-lies.de

On the one hand, the current popularity of "content marketing" contributes to and represents this development. On the other hand, contemporary developments like data-driven performance marketing or marketing automation mean anything but customer centricity.

Against the background of the marketing stages the question arises *if the evolving (digital) services are really as customer-centric as customer- and value-oriented marketing 5.0 supposes.*

In order to answer this question, the text will be structured in the following way. First, marketing intelligence will be presented as service-boom, along with the corresponding marketing paradigms (section II). Subsequently, marketing intelligence will be conceptualized as zombie-marketing (section III). Through this a trade-off will become visible from the perspective of service.

II. Marketing Intelligence as a Boom of Service Marketing

A. Introduction

The term marketing intelligence refers to developing insights obtained from data to optimize marketing decision-making. Data-driven service marketing mainly refers to the use of data in order to inform and optimize the ways through which market activities are carried out. Big data enables digital marketing to analyze digital large (from terabytes to exabytes) and complex (from sensor to social media) data and to react in real time. This is one of the core capabilities of digital marketing techniques. Currently, practices are experiencing the shift from "big data to big impact" [12]. Data-driven marketing evolves from creativity, e.g., designing advertisements, to marketing technology, e.g., block chains to increase digital trust [13].

The digitized world provides lots of data sources, such as e-mail-marketing, clicking behavior on websites, individual and brand community-driven social and mobile media data, cars' sensor data, and many more. This data is likely to increase customer insights and deliver even personalized services in real-time: digital self-services like check-in terminals, digital delivery status services, insurance rates based on driving behavior, alerts to bank customers for unusual account activity, customized mobile adds, car connected sensory maintenance services, new digital interaction channels, as well as conversational commerce are the range of services which directly or indirectly depends on big data growth. Agile service marketing requires at least three dimensions. It means the acceleration of marketing management to provide outcomes faster, by means of agile management processes (e.g., scrum or hackathons applied to marketing). It depends on an agile mindset (i.e., a culture of digitality, which is inspired by the opportunities provided by digital technologies), and it needs the application of agile marketing tools (e.g., automated or real time marketing tools) [14], [15].According to the examples and structure of the services above, the following table shows some data-driven services (see Table I).

In the following, selected data-driven or data-impacted services will be introduced that view marketing intelligence as a drive of a service marketing boom.

B. Marketing Intelligence as a Customer Service Journey

The touchpoints of the customer journey are viewed as a source of big data: Each touchpoint can not only represent a service point, but also provide data for further service optimization. In this way, big data, marketing intelligence and AI provide new opportunities: marketing intelligence seeks to map the touchpoints to understand customer experience and optimize marketing services, as well as customer experience, by providing added values.

Customer experience indicates the perception and evaluation of a customer's interactions with a company across all of its touchpoints – including employees, channels, systems, products and services – connected to the customer lifecycle contribute to the resulting perceptions about the brand [16]. Additionally, touchpoints contribute to the perceived service quality. Product presentations that combine the real product with virtual reality are perceived more positively than those that can only be perceived virtually [17]. Another example are targeting technologies. They are one important data-driven marketing technique to optimize customer experience and, thus, marketing services. Mobile technologies enable marketers to identify and address specific customer communities and/or customers personally. This procedure is known as "mass customization" which refers to group-related or even individualized mass production: the individualized car, the personally branded chocolate, or the personally configured sports shoe [18]. This means that the consequence of micro targeting is micro marketing. In the specified application of geo marketing, micro marketing involves small segments, down to the personal level (segment of one).

C. The Customer Lifecycle as a Service Journey

"While the product lifecycle, which is often contained in conventional marketing textbooks, follows standardized mass marketing, the customer lifecycle emphasizes individual, personalized customer interaction." [19] The marketing concept of lifecycles is derived from biology. The individual lifecycle enables marketing to

TABLE I. Marketing Intelligence as a Service Marketing Boom; Source: own Table

Marketing Service	Marketing Intelligence as a boom of services
Services as added values of products	• The customer journey as customer analytics to derive new services • Big data to enhance customer retention • Big data as provider of ample opportunities to organize services on the customer journey • Voice-based interaction within the (digital) customer journey • Personalized and targeted online and mobile ads • Mobility information for transportation management in urban areas
Services as intangible outcomes increasing customer satisfaction	• The internet, its search engines and data as (inter) corporate services • Recommender systems generate personalized predictions about product liking to increase customer satisfaction by decreased complexity in decisions • The customer journey as provider of interaction opportunities and perceived service quality • The semantic web with speech recognition technology to provide messaging services and virtual assistants • Evolving conversational commerce • Mobile marketing and services as the "mobile revolution"
Service as dominating logic of stakeholder exchange	• Big data to optimize digital customer experience management • Touchpoint management connected to the customer lifecycle in order to enhance brand perception • Viral processes as the outside-in paradigm of marketing

address customers specifically. Digital self-services, e.g., in the process of service delivery or self-service terminals, are adapted to enhance the e-satisfaction of customers [20]. Thus, the idea of customer relationship management (CRM) has entered a "new era" [21] in which it is able to address customers in an automated way, relevant to their personal stage within the lifecycle. Data analysis assists marketing, for instance in acquiring new customers, making existing customers create more profits and maintaining valuable customers [22]. Hence, CRM becomes an analytic approach by tracking data to derive new customer insights from it, and to apply new services. Those technologies may enhance customer centricity. "Gaps between what marketing textbooks prescribe and the real-world confronting marketers need to be narrowed." [23]

One opportunity for lifecycle marketing is retention services. Customer retention means to avoid their defection, aiming at optimizing the lifecycle. One approach for this is the improvement of service quality which requires outside-in thinking or inbound marketing. Instead of interrupting advertising, it is based on interactivity and commitment. This mode became popular as pull marketing. Here, the initiative comes from interested parties [24].

D. Content Marketing as Interaction-Based Services

At first glance, content marketing is not driven by big data as "content" is generated by the sharing, liking, and posting of individual social media users. A second glance reveals that content marketing and big data are closely interlinked by the algorithms of search engines. From a media management point of view, the popularity of social media crucially depends on the ability of search engines which recognize that users contemporarily prefer social media sites to interact with each other, topics, or brands they (dis-)like. Search engines recognize the related web-traffic. Ranking high in search results makes them attractive for social media advertising. Thus, social media platforms are marketing intelligence-driven services [25].

Social media is much more than "media" to perceive content. From a customer's point of view, they provide service rooms for interaction with other users and corporations. Accordingly, they are digital service institutions. Characteristics like contexts, aesthetics, emotions, and symbolic aspects of customer experiences have to be taken into account as soon as they are likely to generate values for the customer [26].

The service management of brands can attempt to leverage social media's connectedness and get consumers to play the brand's game, by creating branded artefacts, social rituals, or cultural icons on behalf of the brand [27]. Hence, social media operates as a foundation for the value fusion of brands. Value fusion is defined as a value that can be achieved by consumers and firms simultaneously, just by being on the mobile network [28]. This fusion of values enlarges and applies the notion of the "value-in-use" idea of services, which supplements the "value-in-exchange" approach, i.e., perceiving values by consumption. Values-in-use emphasize the meaning of added values like services, ideas, information, customer service and payment, as well as invoicing procedures [29]. They provide another source of "creative value creation". Corporations currently depend on digital agile methods to synchronize corporate with societal values [30]. Service as interaction and the convergence of media also lead to converging requirements of service and reputation management.

E. Semantic Marketing as Conversational Services

The Semantic Web represents a revolution for the form of access and storage of information. From a marketing strategic point of view, a crucial change from web 2.0 to web 3.0 is the recognition of "meaning". The benefit of semantics consists in bridging nomenclature and terminological inconsistencies to include underlying meanings in a unified manner [31].

A central popular function of semantic marketing are chatbots (a blending of words derived from "chat" for "chatting" and "bot" for robots). "The goal of chatbots is to have a conversation with humans, so communication with humans is the primary role of chatbots. It is desirable that access to information be as easy as possible for the person and the messaging platforms are selected as convenient platforms for people to use for daily communication." [32] Chatbots operating with artificial intelligence enable marketers to create highly personalized customer experiences. It increases an organization's responsiveness and solves customers' problems.

Chatbots are one example for semantic marketing. They are able to recognize meaning. Artificial intelligence, speech recognition technology, messaging services as well as virtual assistants allow this evolvement of chatbots [33]. Consequently, customer experience becomes enlarged as brands can be perceived through voice-based interaction, which enables personal dialogues within the (digital) customer journey [34]. This development has to be seen as a part of the ongoing developing conversational commerce. This trend entails preparing, accompanying, and executing purchasing processes by (digitized) voice. "Conversational commerce may involve human interaction as well as bots." [35] Absence of conversation is a big barrier to making brand experience become social.

F. Mobile Marketing as Mobile Services

Mobile devices, especially the smartphone, accompanied by the semantic web (web 3.0) enable internet-based services to interact with web users by means of voice recognition. They have also made marketing mobile. Proximity technologies, e.g., near field communications, frame the backbone of mobile touchpoints within the customer journey.

Mobile devices are adapted to enrich this journey: Researchers analyzed the log data in location-based networks to uncover user profiles; these automatically discovered user profiles have the potential to be subsequently applied to location-based targeted marketing. Through analyzing the dynamics of local communities or customer habits, it is possible to predict their changing product/service preferences. Here, one step towards the integration of mobile online and traditional offline marketing, e.g., shopper marketing or in-store marketing, becomes apparent and confirms the idea of servitization. A fusion of values between brands and their products and/or corporations across entire social networks is accelerated. This development sometimes is called "mobile revolution" and indicates a major shift within marketing [36].

G. Geomarketing as Proximity Services

The mobility of marketing itself is not new [37]: From today's perspective, marketing on the radio represents an early form of mobile marketing [38]. In contrast to the early mobile marketing the digitized (geo-) marketing era enables content-specific, individualized, visual and/or conversational as well as geo-located marketing. Smartphones are usually equipped with GPS (Global Positioning System) which enable customized one-to-one marketing. This development led to geomarketing. It uses geographical concepts and instruments, maps, statistics, and information technology. These technologies have opened the opportunity to develop marketing and service marketing strategies and instruments which aim to increase customer satisfaction with local and/or situational marketing measures [39]. Thus, geomarketing makes mobile services become spatial marketing [40].

Recommender systems benefit from mobile services. They provide personalized predictions about product liking by filtering the past behavior and preference statements of consumers. Mobile social media marketing is the digital platform for influencer marketing and a source of big data at the same time. Traditional advertising methods

are shifting toward methods of personalized and targeted online and mobile ads. Another important feature is to apply mobility information for transportation management in urban areas. Marketing began to individually accompany the customers almost everywhere: in the car, at work, in the fitness club, etc. It provides room for geomarketing as creative communication, campaigning and/or services [41].

H. Customized Content as Individual Services

Within internet marketing, online targeting represents the target group-specific display of offers and advertising measures, i.e., customized content. From a service point of view targeting is the basis to provide location-dependent communication, e.g., sports advertising in stadiums or taxi advertising after busses have been delayed and the delay has been digitally recorded online by the local transport companies. This (creative) potential is supported above all by mobile devices. "It is becoming easier to spot market opportunities and solve customer problems more efficiently and cheaply. That's why it's becoming increasingly important to foster disruptive thinking and ignite creativity." [42]

There are many more digital targeting technologies, e.g., retargeting (visitors to an online shop and/or a website are again addressed with advertising on other websites) or photo targeting. Companies such as Pinterest and Snapchat are already designing targeting solutions based on photos posted. Microtargeting approaches analyze consumer variables using statistical techniques and/or machine learning to identify individuals most likely to engage in specific behaviors or respond to marketing in specific ways. For example, marketers use micromarketing to identify customers most likely to respond to tailored advertising campaigns [43]. For example, all credit card users who comment on a certain hashtag on Twitter can be addressed individually. In the specified application of geomarketing, micro marketing involves small segments, right down to the personal level (segment of one).

I. Marketing Intelligence as New Service Paradigm

Looking back, marketing 4.0 and web-based services contemporarily represent a central aspect of service marketing and create a boom of services in marketing intelligence. The broad range of data-driven service opportunities even initiated new business models, e.g., search engines or social media/mobile providers. They contribute to and sustain consumer-based customer management. This implies marketing 3.0, i.e., people-focused thinking. Instead of planning inside-out, the power of viral processes requires the community paradigm of marketing, i.e., the transformation from inside-out to outside-in [44]. All in all, the methodological convergence of marketing and public relations as reputation has to be mentioned, as content marketing means to apply ideas and measures of public relations management (author). Digital service marketing depends on several paradigmatic thoughts, to turn data into customer-centric services: people-focused thinking, agility as well as interactivity are just three crucial thoughts (see Table II). They make digitization become a service culture of "digitality", i.e., a culture shaped by digitization.

In the following, it will be shown that these paradigms are essential for marketing 4.0 to contribute to customer-centric servitization, since the downside of marketing intelligence hasn't been mentioned yet.

III. Data-Driven Zombie Marketing: Marketing Intelligence as Bust of Service Marketing

A. Introduction

Service is not just service. The (digital) evolvement of marketing services have a downside which heavily impacts services. At least since the popularity of so-called "self-services", especially within retail, banking, transportation or hospitality, it goes without saying that marketing services broadly become perverted. The first aim of these services is to reduce costs and not to satisfy customers. Also, the digital call center technologies, established since about the 1990s, were introduced as customer relationship management. Contrarily, these technologies meant the end of the personal customer manager in many sectors, they represent a loss of personal contact. Research finds that customers are less satisfied with the call center services than they are with the more traditional office-based (in-person) services [45]. "While the ways firms interact with customers have changed dramatically, customer' desire for good service have not changed." [46] This leads to some broad research, e.g., in waiting time. "The problem of waiting is important in service activities when customers are passive, often standing in a queue." [47] Vandermerwe and Rada [48] admit that self-services are part of the servitization, but also mean that costs are passed on to the customer. Against this background, Howard and Worboys [49] ask if "self-service" is an oxymoron, i.e. a term that contradicts itself.

Alluding to the rising threat of so-called "zombie corporations", these services are called "zombie services" to distinguish them clearly from customer satisfaction-related services. Zombie companies keep on operating even though they are dead, as they are generating persistent negative equity [50]. They operate at the expense of others, represent a threat by lacking digital customer protection [51] and endanger the trustworthiness of management and brands. These aspects are congenially shared by these "zombie services".

In the following, some insights into these facets of marketing intelligence as performance marketing are detailed, to elaborate the downside of data-driven marketing shaping zombie services (see Table III).

TABLE II. Marketing Intelligence as Service Marketing; Source: own Table

Paradigmatic Service Thoughts	
People-focused/ thinking	Marketing 3.0 as the paradigmatic frame of evolving digital customer-centric marketing services
Agile Thinking	Digitality as a highly dynamic environment requiring steady marketing and branding innovation, e.g., using agile marketing methods like scrum marketing or serving viral processes by social media marketing,
Interactive Thinking	Digitality as an updated understanding of market competition, e.g., crowds as co-designers, crowdsourcing, co-operative thinking
Customized Thinking	Digitality as an opportunity to apply data for personalized communication, products and services
Retention Thinking	Marketing Intelligence to enhance customer retention services
Outside-in Thinking	Digitality as a power to steadily shift demands of stakeholder threatening market positions
Marketing 3.0 Thinking	Digitality as values e.g., serving the manifold social media challenges to prove brands' community compatibility, fashion trends, purpose-led brands

TABLE III. MARKETING INTELLIGENCE A SERVICE MARKETING BUST; SOURCE: OWN TABLE

Zombie Marketing Service	Marketing Intelligence as burst of services
Services as subtracted values of products	• Self-services as oxymorons [73] • Dynamic pricing as common practice on the web [74] • Data-driven marketing automation as annoying marketing [75] • Data-driven marketing as "party crashing" interactive marketing [76] • Data-driven creativity of marketing techniques as creepy marketing [77]
Services as intangible outcomes increasing customer dissatisfaction	• Digital channels as a loss of personal interaction [78] • Migrating customers to online channels as a force creating resistance and customer dissatisfaction [79] • Programmatic native advertisement as deception and misleading [80] • "Customer reviews" as incentives for manipulation [81] • Data algorithms making social media a fertile ground for the virality of fake news [82] • Recommender systems as major culprits of misinformation [83] • Social and mobile media a self-feeding data farms [84]
Service bust as dominating logic of stakeholder exchange	• The notion of marketing automation is its repeatability in which marketers do not have to intervene [85] • Social media as a new era of "information warfare" [86] • Digital key performance indicators with new attention as marketing is coming under pressure to succeed [87] • Marketing practitioners are under increasing pressure to demonstrate their contribution to the firm's performance [88] • Marketing intelligence as lean marketing [89]

B. Digital Self-Services as Zombie Services

At first glance, research regarding e-satisfaction seems to confirm the service boom of digital marketing, as there is a positive relationship between the e-CRM activities on a website and customer satisfaction with the website. A second glance of the same research reveals: "If e-CRM is related to satisfaction, the relationship is not strong. Thus, failure of CRM implementation may not be because the implementation is a failure but because there just is not much that can happen." [52]

"Automated and online migrations present cost savings opportunities as well as risks to customer satisfaction and brand health. More specifically, migrating customers to online channels may create resistance and customer dissatisfaction, as customers may feel forced to use new channels." [53] Research shows that forced use leads to negative attitudes toward using digital services [54]. "The irony is that the more hi-tech our world gets, the more our clients value a personal touch. Our clients don't want group spam — they want to be treated like individuals." [55]

Research asks which factors influence the acceptance of such new technologies: usefulness, ease of use, enjoyment and anxiety are exemplified factors impacting the design of technology acceptance models [56]. Depending on the digital readiness, these services contribute to the e-satisfaction, i.e., consumers' judgment of their digital experiences [57]. The service research mentioned in the previous chapter supposes that established services are appropriate and designed to increase satisfaction.

If research aims at identifying zombie services, it depends on analyzing customers' dissatisfaction: rude employees, poor attitudes, overall poor service, employees socializing, not paying attention to customers as well as slow service are fruitful research findings [58]. Today, the term "service" must be clearly distinguished in customer-oriented services and zombie services, which are developed to reduce costs at the expense of customers. Whether marketing and public release management methodologically converge, this clear distinction is important for the success of management to keep reputation, trust, and customer satisfaction.

Self-services are one popular example for the rising problem of zombie services which are "dead" as they are vitalized by the opposite of customer-centric thinking and service dominant logics. And there are more examples as the following will show.

C. Data-Driven Marketing as "Party Crashers"

"The Web was created not to sell branded products, but to link people together in collective conversational webs. As more branding activity moves online, marketers are confronted with the realization that brands are not always welcome in social media." [59] In this regard, brands and marketing services could be viewed as "party crashers".

Marketing and branding as party crashing zombie services can be exemplified with advertising clutter. This is something already prevalent in offline advertisement and has increased in the digital age. The degree of advertising pressure put on consumers in a particular medium is indicative of advertising clutter. It includes variables like overload, intrusiveness (reactance) and competitiveness (interference) [60]. This clutter is one light version of social media crashing that evolves through the quasi optimization of data-driven advertising.

Programmatic advertising, for example, is one development within data-driven marketing. It applies real time targeting and describes the data supported trading of online advertising spaces. Advertising can be booked fully automatically and adapted situation-specifically for mobile recipients. This digital service means customized advertising on the one hand. On the other hand, it contemporarily leads to annoying, deceptive, and even misleading advertisement. The downside of customized advertising is the risk of being perceived as intrusive [61]. Customized marketing can be perceived as an invasion of privacy, repetition of personalized messages and aggressive tactics which turn creative marketing into "creepy marketing" [62]. Advertisement is adjusted to the personal interests and click behavior of web users. Messages that interrupt a consumer's online activity create feelings of ill will towards the brand [63].

Native advertising is a "try harder attempt" to increase the acceptance of digital ads. It involves "(…) presenting online content to consumers with advertisements that resemble, in format and content, the non-advertising content that is published on the same platform." [64] This advertising is called "native" as it adjusted its look and feel to its "natural environment", e.g., the design of social media posts. Native advertising includes a wide variety of advertising formats, e.g., one-off videos, series of articles, blocks of hyperlinks or social media posts. This type of automated and data-driven advertisement is criticized as deceptive and misleading for customers. It thus contributes to the evolving zombie services.

The basis for marketing intelligence is digital data. As mentioned above, customers provide it along the customer journey. Social and mobile media operate as data farms. "Mobile phones and social media are two examples of big data farms steadily seeded by the users." [65] "Social media have a kind of dual nature: they are public, but often feel private." [66] This automatically leads to questions regarding data protection and the abuse of private data. Is social media intelligence adapted to meet customer needs of private data protection? Block chain technology is discussed to keep data within defined and authorized platforms [67] and will probably shape marketing 5.0 as an era of digital trust to regain reputation.

D. Social Media as (Data-Driven) Fake News

Native as deceptive and misleading marketing seems just to be the tip of the iceberg of "zombie marketing". "The openness and timeliness of social media have largely facilitated the creation and dissemination of misinformation, such as rumor, spam and fake news." [68] With the popularity of the web and especially of social media, a new era of "information warfare" has arrived. Various actors, including state-sponsored ones, are weaponizing information on social networks to run fake news, rumors or clickbaits as campaigns with targeted manipulation of public opinion on a specific topic. The actors include digital bots, political organizations as well as corporations, (paid) activists, "true believers" or "useful idiots" [69]. This digital landscape has provided a fertile ground for fake news to become viral. The algorithm on social media allows accounts to target like-minded individuals based on their browsing and interaction behavior, e.g., clicks, shares or posts [70].

Customer reviews of hotels, restaurants, online shops, etc. are a ubiquitous component of online commerce that impacts customer decisions. "Recommender systems have been pointed as one of the major culprits of misinformation spreading in the digital sphere" [71]. There is an incentive to pollute these reviews, toward promoting one's products or in degrading the ones of competitors. This pollution has been identified as a growing threat to the trustworthiness of online reviews [72]. "Fake it till you make it" [73]. The credibility of reviews is fundamentally undermined when businesses commit review fraud, creating fake reviews for themselves or their competitors. It is estimated that up to one third of all "consumer" reviews on the Internet are fake and, thus, lead to "manufactured opinions" [74]. Machine learning approaches are necessary to make fair recommendations [75].

Congenially, "crowdturfing" evolves. This trend is a counterpart to the opportunities of "crowdsourcing". Crowdturfers leverage human-powered crowdsourcing platforms to spread malicious URLs in social media. The term "crowdturfing" is derived from "astroturf" campaigns which are artificially generated publics [76] and manipulate search engines, ultimately degrading the quality of online information and threatening the usefulness of these systems [77]. Bot generated, artificial consumer reviews and crowdturfing exemplify the increasing significance of digital fake news.

Marketing, e.g., within digital brand management, can interact both directly and indirectly with fake news. In some instances, brands are the victims of fake news. At other times, they are the purveyors. Directly, brands can either finance fake news or be the targets of it. Indirectly, they can be linked via image transfer where either fake news contaminates brands or brands validate fake news. "Searching for greater reach, brands tend to associate themselves with the most popular stories– whether these are true or fake." [78] Research reveals that fake news marketing is likely to increase interest in products [79]. Nevertheless, they belong to the category of zombie marketing as they are artificial, and likely to damage reputation.

E. Performance Marketing as Automated Marketing Cost Optimization

Every day, customers voluntarily generate and provide data by detailing their interest and preference regarding products or services in the public domain, through various channels [80]. They generate data pursuant to their personal stage within their customer lifecycle. Marketing automation uses this data to develop new marketing standards. Marketing automation can be characterized as the methodology by which process design and technology may be harmonized, to enhance both the efficiency and effectiveness of marketing execution [81]. The term "marketing automation" was introduced into the digital age by Little at the 5th Invitational Choice Symposium at UC Berkeley in 2001. The nucleus of marketing automation is an automatic ''customization'' or ''personalization'' of marketing mix activities, applied to the customer´s specific lifecycle. Due to the enhanced relevance of the information provided, it is assumed that customers will show increased involvement and pay more attention to the brand's communication [82].

However, the core value of automation is different. It means the repeatability of (digital) marketing measures in which people do not have to intervene. Head stated as early as the 1960s that new methods of automation would make marketing information available much faster than before and provide data that was not previously available [83]. The broad range of digital marketing technologies indicates that branding as operations is gradually becoming digital. This may enhance customer centricity – or the opposite may occur, as marketing automation processes indicate. "Digital marketing may be facing a black cloud on the horizon. There is mounting concern that consumers find some forms of digital marketing to be intrusive and, thus, annoying" [84]. Marketing automation "[…] does not mean that you sit back and let technology do all the work" [85]. The key to the success of marketing automation is understanding the astute preferences of customers, spotting relevant communication triggers, and converting these into relevant, targeted and timely messages that drive more profitable customer behaviour.

One popular technique in which digital customer data is used are automated price adjustments. Dynamic price adjustment is a popular practice on the web. Amazon is considered a pioneer here. It is assumed that users accept this as long as the offer and service are right [86] and exemplify lacking customer centricity.

F. Marketing Intelligence as Performance Marketing: Digitized Lean Marketing

Contemporary digital zombie services are part of the key performance debate. Automation often implies a strong link to data-driven marketing and performance marketing. Data-driven marketing and marketing automation is closely linked to web analytics [87]. Digital marketing, the collection and analysis of web-based campaigns as well as targeting even in real time, increase the ability to measure the performance of marketing. (Digital) key performance indicators are becoming more and more popular, as marketing budgets increasingly need to be justified [88]. "As pressure for accountability cascades through an organization, every functional group is under scrutiny, and those who cannot quantify their impact on generating satisfactory returns on investment are placed in a vulnerable position" [89]. Research shows that some companies use short-term performance indicators at the expense of measuring long-term factors [90], e.g., regarding campaigns. "If you are going to fail, fail fast" [91]. "Performance marketing in its purest form is purely success-oriented. Successful campaign modules (such as texts, keywords, tools and advertising media) are accelerated and expanded. Less successful ones are optimized and eliminated if the defined goals are still missed"[92]. However, hitherto it is not finally clear what kind of metrics will count. Another danger is that the selected media are so far often too obscure. Moreover, there are ad frauds manipulated by bots which fake impressions.

This performance debate is probably as old as marketing itself and represents the updated lean marketing of the 1990s. Lean management is traced back to automotive industries, especially Toyota, which introduced lean thinking in order to reduce non-value activities. This approach can adequately be applied as lean marketing in the form of continuous improvements to eliminate inefficiency, speed up production cycles and increase professionalism [93].

G. Marketing Intelligence as a Lean Inside-Out Paradigm

In retrospect, several paradigmatic thoughts occur, shaping marketing intelligence as digital performance marketing. Analytic or lean thinking make apparent that performance marketing focuses on the roots of marketing 1.0 with markets as sales institutions (Table IV):

TABLE IV. Marketing Intelligence as Performance Marketing; Source: own Table

Paradigmatic Performance Thoughts	
Performance Thinking	Measuring valuable metrics, e.g., key performance indicators like customer engagement, churn rates or conversion rates
Analytic Thinking	Prioritizing evidence-based marketing measures proved by key performance indicators, e.g., using big data to detect new metrics like ratios "selling sentiment".
Lean Thinking	Eliminating inefficiencies, e.g., long-term marketing and/or branding campaigns serving reputation or image management without direct returns
Process Thinking	Identifying value-creating processes, e.g., the customer journey as the process from perception to action of customers and the identification of the customer-specific lifecycle
Automation Thinking	Establishing repeatable (digital) marketing measures in which people do not have to intervene, e.g., marketing automation solutions
Inside-out Thinking	Conceptualizing, planning and controlling due to the traditional thinking of management analytics
Marketing 1.0 Thinking	Referring to the sales paradigm of marketing and at the same time customer value centricity which is currently limited

The vision of big data alongside the automated customer lifecycle is to systematically eliminate inefficient marketing measures in real time. Ideally, digitized performance marketing 4.0 integrates marketing 1.0 by selling as much as possible, and additionally serves marketing 3.0 by optimizing customer needs through big data analysis. However, the reality is different and leads to the first current incompatibility: Digitization as automated marketing represent cost savings opportunities and risks to customer satisfaction. Besides, it is appropriate to damage brand reputation. Nonetheless, marketing seems to meet its limits when research findings show the annoying impacts of marketing. Hence, the status quo of marketing intelligence research is still considered to be in its infancy and is occasionally even described as "embryonic" [94]. Performance marketing as marketing 3.0 still doesn't perform and seems to persist as marketing 1.0.

IV. Conclusion: Marketing 4.0 as a Trade-Off Between Boom and Bust of Services

In review, service marketing is impacted by many digitized techniques. A broad range of technologies have evolved to increase customer satisfaction, e.g., by customized services. This area of digitization may contribute to service marketing or in fact operate as the opposite when they are primarily designed to reduce costs and/or increase sales. Technologies that contribute to performance marketing by automation on the one hand, and the incentives to pollute the digital landscape by attempting to manufacture opinion on the other hand, lead to "digitality" which seeds zombie services. Zombie services operate at the expense of customer satisfaction. At least since the advent of self-services and call center technologies, it has become popular to pervert the term "service" which is contemporarily a rising threat for (digital) marketing and branding. Against the background of zombie services, marketing 4.0 is likely to disappoint stakeholder expectations and, thus, damage brands. Consequently, the question whether marketing 4.0 means the boom or bust of marketing remains open.

Marketing science and practice need to integrate the elaborated paradigmatic thoughts with the aim of balancing performance requirements and stakeholder values. The digital landscape, shaped by the opportunities of performance marketing and the incentives to pollute the digital environment, require an initiative to foster digital trust which may also impact marketing 5.0. Service marketing needs to embrace a holistic approach, integrating performance marketing, customers, and the claims of other stakeholders [95].

Thus, customer centric marketing 3.0 is a requirement which has still not reached its full application. The need of holistic marketing in order to optimize service and performance is addressed, but marketing intelligence today is lacking in the non-digital and psychological parts of the customer journey. Studies attempt to measure the (non-) financial return on social media investments. However, even the optimistic representatives of "social media return on investment measurement" need to admit: Digital behavior cannot be completely and accurately traced [96]. Marketing intelligence is still in its infancy. Contemporarily, this means that digitized service marketing as a factor of corporate success crucially depends on digitality, i.e., the culturally determined applications of marketing techniques in order to avoid the downside of marketing digitization.

References

[1] P. Kotler, H. Kartajaya, and I. Setiawan, Marketing 5.0: Technology for Humanity, Wiley, Hoboken, 2021.

[2] G. L. Shostack, "Breaking free from product marketing," in *Journal of Marketing*, April 1977, pp. 73-80.

[3] S. Vandermerwe and J. Rada, "Servitization of business: Adding value by adding services," in *European Management Journal*, Vol. 6, Issue 4, Winter, pp. 314-324, 1988.

[4] J. Lies. Die Digitalisierung der Kommunikation im Mittelstand - Auswirkungen von Marketing 4.0, SpringerGabler, Wiesbaden, 2017.

[5] S. Kundu and S.K. Datta, "Impact of trust on the relationship of e-service quality and customer satisfaction," in *EuroMed Journal of Business*, Vol. 10, No. 1, pp. 21-46, 2015, doi: 10.1108/EMJB-10-2013-0053.

[6] P.K. Kannan and A. Li, "Digital marketing: A framework, review and research agenda," in *International Journal of Research in Marketing*, Vol. 34, No. 1, pp. 22–45, 2017.

[7] C. Wood, "Marketing automation: Lessons learnt so far ...," in: *Journal of Direct, Data and Digital Marketing Practice*, Vol. 16, No. 4, pp. 251-254, 2015, doi:10.1057/dddmp.2015.31.

[8] S. L. Vargo and R. Lusch, "From Goods to Service(s): Divergences and Convergences of Logics," in *Industrial Marketing Management*, Vol. 37, No. 3, pp. 254-259, 2008, doi: 10.1016/j.indmarman.2007.07.004.

[9] C. Grönroos, "What can a service logic offer marketing theory?," in *The Service-Dominant Logic of Marketing: Dialog, Debate, and Directions*, R.F. Lusch and S. L. Vargo (Ed.s.) Routledge, London/New York, 2020, pp. 353-364.

[10] S. Vandermerwe and J. Rada, "Servitization of business: Adding value by adding services," in *European Management Journal*, Vol. 6, Issue 4, Winter, pp. 314-324, 1988.

[11] V. Kumar, V. Chattaraman, C. Neghina, B. Skiera, L. Aksoy, A. Buoye, and J. Henseler, "Data-driven services marketing in a connected world," in *Journal of Service Management*, Vol. 24, No. 3, pp. 330-352, 2013, doi: 10.1108/09564231311327021.

[12] H.-c. Chen, R. H. L. Chiang, and V. C. Storey, "Business intelligence and analytics: from big data to big impact," in *Mis Quarterly*, Vol. 36, No. 4, December 2012, pp. 1165–1188, 2012, doi: 10.2307/41703503

[13] D. Shah and B.P.S. Murthi, "Marketing in a data-driven digital world: Implications for the role and scope of marketing," in *Journal of Business Research*, Vol. 125, March 2021, pp. 772-779, 2121, doi: 10.1016/j.jbusres.2020.06.062.

[14] G. Gera, B. Gera, and A. Mishra, "Role of agile marketing in the present era," in *International Journal of Technical Research and Science*, Vol. IV, Issue V, May 2019, pp. 40-44, doi: 10.30780/IJTRS.V04.I05.006.

[15] B. Vassileva, "Marketing 4.0: How Technologies Transform Marketing Organization," in Óbuda *University e-Bulletin*, Vol. 7, No. 1, pp. 47-56, 2017.

[16] D. Nash, D. Armstrong, and M. Robertson, "Customer Experience 2.0: How Data, Technology, and Advanced Analytics are Taking an

Integrated, Seamless Customer Experience to the Next Frontier," in *Journal of Integrated Marketing Communications*, Vol. 1 No. 1, pp. 32-39, 2013.

[17] J. Galán, C. García-García, F. Felip, & M. Contero, "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," *International Journal of Interactive Multimedia and Artificial Intelligence*, Vol. 6, No. 6, pp. 196-207, 2021, doi: 10.9781/ijimai.2021.01.001.

[18] K. N. Lemon and P. C. Verhoef, "Understanding Customer Experience Throughout the Customer Journey," in *Journal of Marketing*, AMA/MSI Special Issue, Vol. 80 (November 2016), pp. 69–96, 2016, doi: 10.1509/jm.15.0420.

[19] E. Sivadasa, R. Grewal, and J. Kellaris, "The Internet as a Micro Marketing Tool: Targeting Consumers through Preferences Revealed in Music Newsgroup Usage," in *Journal of Business Research*, Vol. 41, No. 3, March 1998, pp. 179-186, 1998, doi: 10.1016/S0148-2963(97)00060-X.

[20] S. Boon-itt, "Managing self-service technology service quality to enhance e-satisfaction," in *International Journal of Quality and Service Sciences*, Vol. 7, No. 4, pp. 373-391, 2015, doi: 10.1108/IJQSS-01-2015-0013.

[21] C. Rygielski, J.-C. Wang, and D. C. Yen, "Data mining techniques for customer relationship management," in *Technology in Society*, Vol. 24, No. 4, November 2002, pp. 483-502, 2002, doi: 10.1016/S0160-791X(02)00038-6.

[22] F. Guo and Q. Huilin, "Data Mining Techniques for Customer Relationship Management," in *Journal of Physics: Conference Series*, 910(1):012021, 2017, doi :10.1088/1742-6596/910/1/012021.

[23] E. Gummesson, "Customer centricity: reality or a wild goose chase?," in *European Business Review*, Vol. 20, No. 4, 2008, pp. 315-330, doi 10.1108/09555340810886594.

[24] W. Kamakura, C. F. Mela, A. Ansari, A. Bodapati, P. Fader, R. Iyengar, P. Naik, S. Neslin, B. Sun, P. C. Verhoef, M. Wedel, and R. Wilcox, "Choice Models and Customer Relationship Management," in *Marketing Letters*, December 2005, Vol. 16, Issue 3–4, pp. 279-291, 2005, doi: 10.1007/s11002-005-5892-2.

[25] J. Lies, "Digital marketing: Incompatibilities between performance marketing and marketing creativity," in *Journal of Digital & Social Media Marketing*, Vol. 8, No. 4, pp. 376-386, 2021.

[26] B. Hultén, "Sensory marketing: the multi-sensory brand-experience concept," in *European Business Review*, Vol. 23, Issue: 3, pp. 256-273, 2011, doi: 10.1108/09555341111130245.

[27] S. Gensler, F. Völckner, Y. Liu-Thompkins, and C. Wiertz, "Managing Brands in the Social Media Environment," in *Journal of Interactive Marketing*, Vol. 27, Issue 4, pp. 242-256, 2013, doi: 10.1016/j.intmar.2013.09.004.

[28] B. Larivière, H. Joosten, E.C. Malthouse, M. v. Birgelen, P. Aksoy, W.H. Kunz, and M.-H. Huang, "Value fusion: The blending of consumer and firm value in the distinct context of mobile technologies and social media," in *Journal of Service Management*, Vol. 24, Issue: 3, pp. 268-293, 2013, doi: 10.1108/09564231311326996.

[29] J. Rowley, "Understanding digital content marketing," in *Journal of Marketing Management*, Vol. 24, No. 5-6, pp. 517-540, 2008, doi: 10.1362/026725708X325977.

[30] L. de Chernatony and S. Cottam, "Interactions between organisational cultures and corporate brands," in *Journal of Product & Brand Management*, 17/1, pp. 13–24, 2008, doi: 10.1108/10610420810856477.

[31] A. García-Crespo, R. Colomo-Palacios, J.M. Gómez-Berbís, and F. Paniagua Martín, "Customer Relationship Management in Social and Semantic Web Environments," in *International Journal of Customer Relationship Marketing and Management*, Vol. 1, No. 2, pp. 1-10, 2010, doi 10.4018/jcrmm.2010040101.

[32] U. Arsenijevic and M. Jovic, "Artificial Intelligence Marketing: Chatbots," in *International Conference on Artificial Intelligence: Applications and Innovations (IC-AIAI)*, Belgrade, Serbia, pp. 19-22, 2019, doi: 10.1109/IC-AIAI48757.2019.00010.

[33] S. Tuzovic and S. Paluch, "Conversational Commerce – A new area for Service Business Development," in: *Service Business Development: Strategien – Innovationen – Geschäftsmodelle*; M. Bruhn and K. Hadwich, K. (Ed.s.), Forum Dienstleistungsmanagement, Band 1, SpringerGabler, 2018, Wiesbaden, pp. 81-100, doi: 10.1007/978-3-658-22426-4.

[34] V. Wanick, A. Ranchhod, and C. Gurau, "Digital Interactions and Brand Experience Design: a future perspective, in *Design Management Academy Conference 2017: Research Perspectives on Creative Intersections*, Hong Kong, China, 07-09 Jun 2017., pp. 1263-1281, 2017, doi: 10.21606/dma.2017.129.

[35] N. Piyush, T. Choudhury, and P. Kumar, "Conversational Commerce a New Era of E-Business," in *Proceedings of the SMART - 2016, IEEE Conference ID: 39669 5th International Conference on System Modeling & Advancement in Research Trends*, 25th-27th November, 2016, College of Computing Sciences & Information Technology, Teerthanker Mahaveer University, Moradabad, India, pp. 322-327, doi: 10.1109/SYSMART.2016.7894543.

[36] S. Fan, R. Y.K. Lau, and J. L. Zhao, "Demystifying Big Data Analytics for Business Intelligence Through the Lens of Marketing Mix," in *Big Data Research*, 2 (2015), pp. 28–32, 2015, doi: 10.1016/j.bdr.2015.02.006.

[37] Y.B.B. Öztaş, "The Increasing Importance of Mobile Marketing in the Light of the Improvement of Mobile Phones, Confronted Problems Encountered in Practice, Solution Offers and Expectations," in *World Conference on Technology, Innovation and Entrepreneurship, Procedia - Social and Behavioral Sciences*, 195, pp. 1066-1073, 2015, doi: 10.1016/j.sbspro.2015.06.150.

[38] S. Banerjee and R.R. Dholakia, "Mobile Advertising: Does Location Based Advertising Work?," in *International Journal of mobile Marketing*, December 2008, Vol. 3, No. 2, pp. 68–74, 2008.

[39] M. Cavallone, F. Magno, and A. Zucchi, "Improving service quality in healthcare organisations through geomarketing statistical tools", in *The TQM Journal*, Vol. 29 No. 5, pp. 690-704, 2017, doi: 10.1108/TQM-12-2016-0104

[40] G. Cliquet, "Spatial Marketing", in *Geomarketing – Methods and Strategies in Spatial Marketing*, G. Cliquet (ed.), Geographical Information Systems Series, Iste, 2006, London.

[41] E. Ahmed, I. Yaqoob, I. A. T. Hashem, J. Shuja, M. Imran, N. Guizani, and S. T. Bakhsh, "Recent Advances and Challenges in Mobile Big Data," in *IEEE Communications Magazine*, Vol. 56, No. 2, pp. 102-108, Feb. 2018, doi: 10.1109/MCOM.2018.1700294.

[42] K. N. Lemon, „Die Kunst zum richtigen Zeitpunkt attraktive Kundenerlebnisse zu gestalten: Marketingfähigkeiten für die digitale Welt," in *GfK Marketing Intelligence Review*, Vol. 8, No. 2 (November), pp. 44-49, 2016.

[43] A. L. Metcalf, J. W. Angle, C. N. Phelan, B. A. Muth, and J. C. Finley, "More "Bank" for the Buck: Microtargeting and Normative Appeals to Increase Social Marketing Efficiency," in *Social Marketing Quarterly* 2019, Vol. 25, No. 1, pp. 26-39, doi: 10.1177/1524500418818063.

[44] S. Quinton, "The community brand paradigm: A response to brand management's dilemma in the digital era," in *Journal of Marketing Management*, May 2013, Vol. 29, No. 7-9, pp. 912-932, 2013, DOI: 10.1080/0267257X.2012.729072.

[45] L. Bennington, J. Cummane, and P. Conn, "Customer satisfaction and call centers: an Australian study," in *International Journal of Service Industry Management*, Vol. 11, No. 2, pp. 162-173, 2000, doi: 10.1108/09564230010323723.

[46] M. J. Bitner, "Service and technology: opportunities and paradoxes," in *Managing Service Quality: An International Journal*, Vol. 11, No. 6, pp. 375-379, 2001.

[47] A. Durrande-Moreau, "Waiting for service: ten years of empirical research," in *International Journal of Service Industry Management*, Vol. 10, No. 2, pp. 171-194, 1999. doi: 10.1108/09564239910264334.

[48] S. Vandermerwe and J. Rada, "Servitization of business: Adding value by adding services," in *European Management Journal*, Vol. 6, Issue 4, Winter, pp. 314-324, 1988.

[49] M. Howard and C. Worboys, "Self-service – a contradiction in terms or customer-led choice?," in *Journal of Consumer Behaviour*, Vol. 2, Issue 4, pp. 382-392, 2006, doi: 10.1002/cb.115

[50] S. Urionabarrenetxea, J.D. Garcia-Merino, L. San-Jose, and J. L. Retolaza, "Living with zombie companies: Do we know where the threat lies?," in *European Management Journal*, Volume 36, Issue 3, pp. 408-420, 2018, DOI: 10.1016/j.emj.2017.05.005.

[51] T. Tajti, "Unprotected Consumers in the Digital Age: The Consumer-Creditors in Bankrupt, Abandoned, Defunct and of Zombie Companies," in *Tilburg Law Review*, Vol. 24, No. 1, pp. 3-26, 2019, doi: 10.5334/tilr.139.

[52] R. Feinberg and R. Kadam, "E-CRM Web service attributes as determinants

of customer satisfaction with retail Web sites," in *International Journal of Service Industry Management*, Vol. 13, No. 5, pp. 432-451, 2002, doi: 10.1108/09564230210447922.

[53] P.S.H. Leeflang, P. C. Verhoef, P. Dahlström, and T. Freundt, "Challenges and solutions for marketing in a digital era," in *European Management Journal*, Vol. 32, Issue 1, pp. 1–12, 2014, doi: 10.1016/j.emj.2013.12.001.

[54] M.J. Reinders, P.A. Dabholkar, and R.T. Frambach, "Consequences of Forcing Consumers to Use Technology-Based Self-Service," in *Journal of Service Research*, Vol. 11, No. 2, November, pp. 107-123, 2008, doi: 10.1177/1094670508324297.

[55] C. Wood, "Marketing automation: Lessons learnt so far …," in: *Journal of Direct, Data and Digital Marketing Practice*, Vol. 16, No. 4, pp. 251-254, 2015, doi:10.1057/dddmp.2015.31.

[56] M Blut, C. Wang, and K. Schoefer, "Factors Influencing the Acceptance of Self-Service Technologies: A Meta-Analysis," in *Journal of Service Research*, Vol. 19, No. 4, pp. 396-416, 2016, doi: 10.1177/1094670516662352.

[57] S. Boon-itt, "Managing self-service technology service quality to enhance e-satisfaction," in *International Journal of Quality and Service Sciences*, Vol. 7, No. 4, pp. 373-391, 2015, doi: 10.1108/IJQSS-01-2015-0013.

[58] M.M. Helms and D.T. Mayo, "Assessing poor quality service: perceptions of customer service representative," in *Managing Service Quality: An International Journal*, Vol. 18, No. 6, pp. 610-622, 2008, doi: 10.1108/09604520810920095.

[59] S. Fournier and J. Avery, "The uninvited brand," in *Business Horizons*, Vol. 54, Issue 3, pp. 193-207, 2011, doi: 10.1016/j.bushor.2011.01.001.

[60] F. Rejón-Guardia and F.J. Martínez-López, "Online Advertising Intrusiveness and Consumers' Avoidance Behaviors," in *Handbook of Strategic e-Business Management*, F.J. Martínez-López (Ed.), Springer, Berlin/Heidelberg, pp. 565-586, doi: 10.1007/978-3-642-39747-9_23.

[61] A. Goldfarb and C. Tucker, "Online Display Advertising: Targeting and Obtrusiveness," in *Marketing Science*, Vol. 30, Issue 3, May-June 2011, pp. 389-564, 2010, doi: 10.1287/mksc.1100.0583.

[62] R.S. Moore, M.L. Moore, K. Shanahan, A. Horky, and B. Mack, "Creepy Marketing: Three dimensions of perceived excessive online privacy violation," in *Marketing Management Journal*, Spring 2015, Vol. 25, Issue 1, pp. 42-53, 2015.

[63] K.T. Smith, "Digital marketing strategies that Millennials find appealing, motivating, or just annoying," in *Journal of Strategic Marketing*, Vol. 19, No.:6, pp. 489-499, 2011, doi: 10.1080/0965254X.2011.581383.

[64] B.W. Wojdynski, "Native advertising: Engagement, deception, and implications for theory," in *The New Advertising: Branding, Content and Consumer Relationships in a Data-Driven Social Media Era*, R.E. Brown, V.K. Jones, and M. Wang (Ed.s.), Santa Barbara, CA: Praeger/ABC Clio, pp. 203-236, 2016.

[65] J. Lies, "Marketing Intelligence and Big Data: Digital Marketing Techniques on their Way to Becoming Social Engineering Techniques in Marketing" in *International Journal of Interactive Multimedia and Artificial Intelligence*, Vol. 5, No. 5, pp. 134-144, 2019, doi: 10.9781/ijimai.2019.05.002.

[66] K.V. Rønn and S. Obelitz Søe, "Is social media intelligence private? Privacy in public and the nature of social media intelligence," in *Intelligence and Security*, Vol. 34, No. 3, pp. 362-378, 2019, doi: 10.1080/02684527.2019.1553701.

[67] A. Boukis, "Exploring the implications of blockchain technology for brand–consumer relationships: a future research agenda," in *Journal of Product and Brand Management*, Vol. 29 No. 3, pp. 307-320, 2019, doi: 10.1108/JPBM-03-2018-1780.

[68] L. Wu, F. Morstatter, K.M. Carley, and H. Liu, "Misinformation in Social Media: Definition, Manipulation, and Detection," in *ACM SIGKDD Explorations Newsletter*, vol. 21, no. 2, November 2019, pp. 80-90, 2019, doi: 10.1145/3373464.3373475.

[69] S. Zannettou, M. Sirivianos, J. Blackburn, and N. Kourtellis, "The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans," in *ACM Journal of Data and Information Quality*, Vol. 11, No. 3, Article 10, pp. 10.1-10.37, 2019, doi: 10.1145/3309699.

[70] Z.F. Chen and Y. Cheng, "Consumer response to fake news about brands on social media: the effects of self-efficacy, media trust, and persuasion knowledge on brand trust," in *Journal of Product and Brand Management*, Vol. 29, No. 2, pp. 188-198, 2019, doi: 10.1108/JPBM-12-2018-2145.

[71] M. Fernandez and A. Bellogin, "Recommender Systems and Misinformation: The Problem or the Solution?," in *OHARS Workshop. 14th ACM Conference on Recommender Systems*, pp. 22-26, Sep 2020, [Online].

[72] A. Fayazi, K. Lee, J. Caverlee, and A. Squicciarini, "Uncovering Crowdsourced Manipulation of Online Reviews," in *Development in Information*, August 2015, pp. 233–242, 2015, doi: 10.1145/2766462.2767742.

[73] M. Luca and G. Zervas, "Fake it till you make it: reputation, competition, and Yelp review fraud," in *Management Science*, Vol. 62, No.12, pp. 3412-3427, 2016, doi: 10.1287/mnsc.2015.2304.

[74] M. Zhuang, G. Cui, and L. Peng, "Manufactured opinions: The effect of manipulating online product reviews," in *Journal of Business Research*, Vol. 87, pp. 24-35, 2018, doi: 10.1016/j.jbusres.2018.02.016.

[75] J. Bobadilla, R-Lara-Cabrera, Á. González-Prieto, and F. Ortega, "DeepFair: Deep Learning for Improving Fairness in Recommender Systems," *International Journal of Interactive Multimedia and Artificial Intelligence*, Vol. 6, Nº6, pp. 86-94, 2020, doi 10.9781/ijimai.2020.11.001.

[76] J. Song, S. Lee, and J. Kim, "CrowdTarget: Target-based Detection of Crowdturfing in Online Social Networks," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (CCS '15). Association for Computing Machinery, New York, NY, USA, pp. 793–804, doi: 10.1145/2810103.2813661.

[77] K. Lee, P. Tamilarasan, and J. Caverlee, "Crowdturfers, Campaigns, and Social Media:Tracking and Revealing Crowdsourced Manipulation of Social Media," in *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, Vol. 7, No. 1, pp. 331-340, 2013, available: https://ojs.aaai.org/index.php/ICWSM/article/view/14384.

[78] P. Berthon, E. Treen, and L. Pitt, "How Truthiness, Fake News and Post-Fact Endanger Brands and What to Do About It," in *NIM Marketing Intelligence Review*, Vol. 10, No. 1, pp. 19-23, 2018, doi: 10.2478/gfkmir-2018-0003.

[79] A. Rao, "Deceptive Claims using Fake News Marketing: The Impact on Consumers", 2018, doi: 10.2139/ssrn.3248770.

[80] M. Anshari, M.N. Almunawar, S.A. Lim, and A. Al-Mudimigh, "Customer relationship management and big data enabled: Personalization and customization of services," in *Applied Computing and Informatics*, Vol. 15, Issue 2, July 2019, pp. 94-101, 2018, doi: 10.1016/j.aci.2018.05.004.

[81] B. Biegel, "The current view and outlook for the future of marketing automation," in *Journal of Direct, Data and Digital Marketing Practice*, 10, pp. 201-213, 2009, doi: 10.1057/dddmp.2008.37.

[82] I. Heimbach, D.S. Kostyra, and O. Hinz, "Marketing Automation," in *Business and Information Systems Engineering*, Vol. 57, No. 2, pp. 129–133, 2015, doi: 10.1007/s12599-015-0370-8.

[83] G.W. Head, "What does automation mean to the Marketing man?," in *European Journal of Marketing*, Vol. 24, No. 4, pp. 35-37, 1960.

[84] K.T. Smith, "Digital marketing strategies that Millennials find appealing, motivating, or just annoying," in *Journal of Strategic Marketing*, Vol. 19, No.:6, pp. 489-499, 2011, doi: 10.1080/0965254X.2011.581383.

[85] A. Bagshaw, "Opinion Piece What is marketing automation?," in *Journal of Direct, Data and Digital Marketing Practice*, 17, pp. 84–85, 2015, doi:10.1057/dddmp.2015.46.

[86] R. Kalka and A. Krämer, „Dynamic Pricing: Verspielt Amazon das Vertrauen seiner Kunden?," in *Absatzwirtschaft*, 16.02.2016, [Online] Available: http://www.absatzwirtschaft.de/dynamic-pricing-verspielt-amazon-das-vertrauen-seiner-kunden-75271/

[87] J. Järvinen and H. Karjaluoto, "The use of Web analytics for digital marketing performance measurement," in *Industrial Marketing Management*, Vol. 50, pp. 117-127, 2015, doi: 10.1016/j.indmarman.2015.04.009.

[88] D.M. Hanssens and K. H. Pauwels, "Demonstrating the value of Marketing" in *Journal of Marketing*, Vol. 80, No. 6, pp. 173-190, 2016, doi: 10.1509/jm.15.0417.

[89] S. Chiu and T. Domingo, "*Data Mining and Market Intelligence for Optimal Marketing Returns*", Butterworth-Heineman/Elsevier, Oxford/Burlington, 2008.

[90] J. Järvinen , "*The Use of Digital Analytics for Measuring and Optimizing Digital Marketing Performance,*" Jyväskylä University School of Business and Economics, University Library of Jyväskylä, 2016, [Online]. Available: https://jyx.jyu.fi/bitstream/handle/123456789/51512/978-951-39-6777-2_vaitos21102016.pdf?sequence=1.

[91] M. Jeffery, "*Data-Driven Marketing: The 15 Metrics Everyone in Marketing Should Know*," Wiley and Sons, Hoboken, 2010.

[92] E. Lammenett, „*Praxiswissen Online-Marketing, Affiliate- und E-Mail-*

Marketing, Suchmaschinen-Marketing, Online-Werbung, Social Media, Online-PR", SpringerGabler, Wiesbaden, 2015.

[93] R. Dewell, "The dawn of Lean marketing," in *Journal of Digital Asset Management*, Vol. 3, No. 1, pp. 23-28, 2007, doi: 10.1057/palgrave.dam.3650054.

[94] A. Amado, P. Cortez, P. Rita, and S. Moro, "Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis," in *European Research on Management and Business Economics*, Vol. 24, Issue 1, January–April 2018, pp. 1-7, 2018, doi: 10.1016/j.iedeen.2017.06.002

[95] N. Blackburn, V. Hooper, R. Abratt, and J. Brown, "Stakeholder engagement in corporate reporting: towards building a strong reputation," in *Marketing Intelligence and Planning*, Vol. 36, No. 4, pp. 484-497, 2018, doi: 10.1108/MIP-10-2017-0236.

[96] D.L. Hoffman and M. Fodor, „Can you measure the ROI of your social media marketing?," in *MITSloan Management Review*, Fall 2010, Vol 52, No. 1, pp. 41-49, 2012.

Jan Lies

Prof. Dr. Jan Lies has a doctorate and postdoctoral habilitation in economics. Since 2013 he has been professor for business administration, with a focus on corporate communications and marketing at FOM University of Applied Science in Dortmund, Germany. His research involves evolutionary and behavioral economics, as well as digital marketing, PR-management and change communications. Digital marketing is one of his research areas, which demonstrates the heavy impact of evolutionary processes on corporate success.

LA UNIVERSIDAD EN INTERNET

Rectorado
Avenida de la Paz, 137
26006 Logroño (La Rioja)
t (+34) 941 21 02 11

www.unir.net