A Robust Framework for Speech Emotion Recognition Using Attention Based Convolutional Peephole LSTM

Ramya Paramasivam[1], K. Lavanya[2], Parameshachari Bidare Divakarachari[3]*, David Camacho[4]

[1] Department of Computer Science and Engineering, Mahendra Engineering College (Autonomous), Mallasamudram (India)
[2] Department of Electronics and Communication Engineering, Velammal Engineering College, Chennai (India)
[3] Department of Electronics and Communication Engineering, Nitte Meenakshi Institute of Technology, Bengaluru (India)
[4] Departamento de Sistemas Informáticos, Universidad Politécnica de Madrid, Calle Alan Turing s/n, Madrid, 28031, Madrid (Spain)

* Corresponding author: paramesh@nmit.ac.in

Received 13 May 2024 | Accepted 27 September 2024 | Early Access 5 February 2025

Abstract

Speech Emotion Recognition (SER) plays an important role in emotional computing which is widely utilized in various applications related to medical, entertainment and so on. The emotional understanding improvises the user machine interaction with a better responsive nature. The issues faced during SER are existence of relevant features and increased complexity while analyzing of huge datasets. Therefore, this research introduces a well-organized framework by introducing Improved Jellyfish Optimization Algorithm (IJOA) for feature selection, and classification is performed using Convolutional Peephole Long Short-Term Memory (CP-LSTM) with attention mechanism. The raw data acquisition takes place using five datasets namely, EMO-DB, IEMOCAP, RAVDESS, Surrey Audio-Visual Expressed Emotion (SAVEE) and Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D). The undesired partitions are removed from the audio signal during pre-processing and fed into phase of feature extraction using IJOA. Finally, CP-LSTM with attention mechanisms is used for emotion classification. As the final stage, classification takes place using CP-LSTM with attention mechanisms. Experimental outcome clearly shows that the proposed CP-LSTM with attention mechanism is more efficient than existing DNN-DHO, DH-AS, D-CNN, CEOAS methods in terms of accuracy. The classification accuracy of the proposed CP-LSTM with attention mechanism for EMO-DB, IEMOCAP, RAVDESS and SAVEE datasets are 99.59%, 99.88%, 99.54% and 98.89%, which is comparably higher than other existing techniques.

Keywords

Attention Mechanism, Convolutional Peephole Long Short-Term Memory, Feature Selection, Improved Jellyfish Optimization Algorithm, Speech Emotion Recognition.

DOI: 10.9781/ijimai.2025.02.002

I. Introduction

THE advent in the era of artificial Intelligence has attracted a greater number of researches to work on human-computer interaction [1]. The affective computing plays a significant role in interaction among human and the computer which is endowed in computers with the ability to observe and exhibit the emotion of the humans [2]. In general, the emotional state of the humans is evaluated based on the speech, body language and their facial expression. Among these, speech is a kind of natural method utilized for human communications which is comprised with linguistic and paralinguistic information [3]-[5]. The information related to context and language is present in the linguistic information whereas the paralinguistic information has information of gender, age, emotions and some more unique attributes [6]. Several researches reveals that the audio signal acts as a simple mean to execute the link among human and computers which in turn become familiar with human voice and helps to predict the emotion [7]. An affective Speech Emotion Recognition (SER) is characterized by features on the basis of speech signals such as bandwidth, duration and frequency. Automatic approach involved in SER helps in various real time applications based on recognizing and detecting the mental and emotional state of individuals [8], [9]. SER is vastly utilized in real time applications such as human-computer interaction, call centers, healthcare and automated translation systems.

Speech emotion recognition has received a great deal of attention in Psychology and cognitive science, but data science has recently contributed significantly to the advancement of SER by highlighting a particularly captivating and motivating feature of human-machine interaction in voice communication. Further, it can be applied to the field of e-learning, automobile board systems, autonomous remote call centers, and student emotions recognition during lectures. This motivates the researchers to work on a few well-known voice computation and classification techniques to extract sentiments from

Please cite this article as: R. Paramasivam, K. Lavanya, P. B. Divakarachari, D. Camacho. A Robust Framework for Speech Emotion Recognition Using Attention Based Convolutional Peephole LSTM, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 4, pp. 45-58, 2025, http://dx.doi.org/10.9781/ijimai.2025.02.002

audio inputs, utilizing deep learning techniques like audio signal preprocessing, feature extraction and selection approaches, and eventually determining the accuracy of the suitable classifier. The process of recognizing the human emotion is complex process due to the dependency on various factors such as speaker, gender, age and dialect. In every individual stage of SER, the data processing, feature extraction, feature selection and classification plays a major role [10]-[12]. The stage of pre-processing is based on normalizing the signals, removal of noise and artifacts. The feature extraction helps to mine out the salient features of emotion using different feature extraction techniques. Next to this, feature selection takes place which has a great role in reducing the complexities of SER and finally, the classification is performed with the help of machine learning/deep learning techniques [13], [14]. Additionally, most of the datasets based on speech emotion are comprised only with utterance level label classes that hold the information of the emotion [15]-[17]. The researches have utilized different machine learning and deep learning techniques for an effective SER, but the usage of deep learning architectures exhibits better results in modelling emotions [18], [19]. The usage of deep learning techniques provides promising results, having the ability to define the low-quality attributes to higher attributes. Moreover, they have tendency to handle the unlabeled data, manage with complex speech attributes and processing large datasets [20]. By considering this, this research focuses on introducing an optimization-based feature selection approach and an effective recognition of emotion using a novel deep learning approach.

A. Contribution

The major contribution of this research study is listed as follows:

1. The acquisitioned raw data is pre-processed and fed into the stage of feature extraction where the prosodic features and the acoustic features are extracted.
2. The feature selection performed using the proposed Improved Jellyfish Optimization Algorithm (IJOA) by introducing sine and cosine factors at stage of exploration and premature convergence strategy is utilized in the stage of exploitation.
3. Finally, the SER takes place with the help of the Convolutional Peephole Long Short-Term Memory (CP-LSTM) with attention mechanism. The combination of the proposed CP-LSTM with attention mechanism helps in an effective SER by recognizing the signal pattern and results in better classification accuracy.

The remainder of the manuscript is structured in the following manner: Section II presents details about the recent research based on SER and Section III presents the overall process involved in the proposed framework while recognizing the human emotions; Section IV presents the experimental outcome while evaluating CP-LSTM and finally, overall conclusion of this research is described in Section V of the manuscript.

II. Related Work

In this section, the recent researches based on speech emotion recognition are discussed along with their advantages and drawbacks.

Yildirim et al. [21] introduced a modified feature selection approach for speech emotion recognition. This research utilized cuckoo search algorithm along with the non-dominated sorting genetic algorithm. The generation phase of the initial population is modified and the feature selection is performed by maximizing the classifier accuracy and minimizing the number of features using binary cuckoo and non-dominated sorting genetic algorithm. At last, the results are evaluated based on machine learning techniques. However, performing an exhaustive search for all subset space is infeasible.

Agarwal and Om [22] introduced the optimized Deep Neural Network (DNN) for speech emotion recognition. The speech signals are de-noised using Adaptive Wavelet Transform with Modified Galactic Swarm Optimization (AWT-MGSO). The de-noised output is provided to the phase extracting the features and the feature selection is performed using Adaptive Sunflower Optimization (ASFO). Finally, classification is performed using DNN-Deer Hunting Optimization (DHO). However, implementing a greater number of optimization techniques results in computational complexity.

Manohar and Logashanmugam [23] developed a hybrid deep learning approach along with feature selection for recognizing the emotion of speech using Deer Hunting with Adaptive Search (DH-AS). The acquisitioned data is pre-processed by median filtering and artifact removal, then it is subjected to the stage of feature extraction. After this, the selection of optimal features takes place using DH-AS and the classification is performed using hybrid DNN and Recurrent Neural Network (RNN). However, the suggested approach was suitable only for the balanced dataset and was complex due to its recurrent nature.

Mustaqeem and Kwon [24] developed an optimal feature selection approach using two stream Deep Convolutional Neural Network (D-CNN). The spectrum and the spectrogram of the speech signals are considered, then the high-level discriminative features are obtained using 2D and 1D CNN. The Iterative Neighborhood Component Analysis (INCA) was utilized in the process of selecting the optimal features by removing the redundant information. Finally, the classification was performed using a softmax layer. However, the improper pre-processing leads to discrepancies and redundancies that diminish the efficiency of the overall model.

Chattopadhyay et al. [25] introduced a hybrid feature selection approach using Clustering based Equilibrium Optimizer and Atom Search Optimization (CEOAS) for recognizing emotions from speech signals. The features such as Linear Prediction Coding (LPC) and Linear Predictive Cepstral Co-efficient (LPCC) were extracted from audio signals. At last, the results are evaluated with two classifiers like K-Nearest neighbor and Support Vector Machine (SVM). The suggested approach diminishes the feature dimension and helps to enhance the classification ability. However, CEOAS exhibits premature convergence at the time of iterative searching process.

Kanwal et al. [26] developed a Density based Spatial Clustering with Noise Genetic Algorithm (DGA) to recognize the type of emotion from the speech. The obtained data is pre-processed by removing the unvoiced audio segments and the optimization of feature was performed using DGA. After this, the reduction of features takes place using Principal Component Analysis (PCA) and finally, the classification was performed using Support Vector Machine (SVM). However, the PCA was computationally imprecise for huge datasets.

Barsainyan and Singh [27] introduced optimized speech emotion recognition using 1D CNN. The obtained data is augmented using glottal inverse filtering, silent elimination and noise addition. After this, the feature selection based on spectral contrast, zero crossing rate and amplitude energy are considered. Finally, the classification was performed using normalized CNN and XGB algorithm. However, the suggested framework was not vulnerable to higher datasets due to limited model training.

Sun et al. [28] introduced a speech emotion recognition approach using Improved Masking based Empirical Model Decomposition and Convolutional Recurrent Neural Network (IMEMD-CRNN). Initially, the decomposition of speech was performed using IMEMD, which is based on disturbance assisted EMD that determines the nature of signals. After this, the 43-dimensional time frequency features are used to characterize emotion and the acquired features are fed into CRNN to recognize the emotions. However, the mode mixing occurs which diminishes the capability of IMEMD to decompose the signal.

Ottoni et al. [29] introduced a deep learning approach based on CNN and Long Short-Term Memory (LSTM) which is used to recognize the speech emotions based on meta learning approach. After the stage of data acquisition, the optimizers such as Adam optimizer, Stochastic Gradient and Adagrad were used to select the optimal learning rate for the dataset. Then, data augmentation is performed to enhance the diversity of audio samples. The augmented output is provided to the phase of extracting features and the extracted features are fed into classification which is performed using CNN-LSTM. However, the lack of feature selection leads to computational complexity and diminishes the efficiency of classification.

Nhat Truong Pham et al. [30] proposed Hybrid Data Augmentation (HDA) and Modified Attention-Based Dilated Convolutional and Recurrent Neural Network (mADCRNN) methods for speech emotion recognition. The mADCRNN model learns and extracts utterance-level features from 3D log MelSpec low-level data by combining dilated CNNs and dilated LSTM models with an attention mechanism. While dilated recurrent neural networks solve complex dependencies and the vanishing and inflating gradient problems, dilated CNNs gain wider receptive fields. In addition, the loss functions are rearranged to identify different emotional states by merging the SoftMax loss and the center-based losses. However, the number of layers and parameters in the suggested mADCRNN model made it complicated.

Zengzhao Chen et al. [31] demonstrated a parallel network for multi-scale SER based on connection attention mechanism (AMSNet) for speech emotion recognition. In the meantime, AMSNet enhances feature characterization and feature enrichment by using several speech emotion feature extraction modules based on the temporal and spatial properties of speech signals. varying types of features are given varying weight values by the network fusion connection attention technique. The model's capacity to recognize emotions has increased as a result of the integration of different features through the use of weight values. However, because of their less conspicuous characteristics, neutral emotions are less often recognized. Since the characteristics of neutral emotions were not sufficiently evident and it was challenging to categorize the corresponding voice signal, this problem also occurred here.

Mustaqeem Khan et al. [32] demonstrated a Multimodal Speech Emotion Recognition (MSER) system to effectively identify the speech emotions. The suggested model makes use of both text and audio to accurately predict the emotion label. The suggested model uses CNN to process the text and raw speech signal before feeding the results to appropriate encoders for the extraction of semantic and discriminative features. In order to improve text and auditory cues interaction, the cross-attention mechanism has been implemented to both elements. This allows crossway to extract the most pertinent information for emotion recognition. Eventually, the deep feature fusion technique allows for interaction between various layers and routes by merging the region-wise weights from both encoders. Yet, in order to address the restriction of the complicated interaction between speech signals and transcripts, these investigations disregarded the limitations of past knowledge. Francesco Ardan Dal Ri et al. [33] demonstrated an extensive validation using CNN for speech emotion recognition. Here, the suggested CNN integrated with a Convolutional Attention Block was tested in a sequence of experimentations including a collection of four datasets namely RAVDESS, TESS, CREMA-D, and IEMOCAP. After analyzing the datasets, they executed a cross-validation among emotional classes belonging to every specified dataset, through the purpose to examine the generalization capabilities of extracted features. Apart from the accuracy improvement once it was trained, the minimum accuracies only attained by testing validates the individuality of the individual models on the feature extraction.

Heuristic Multimodal Real-Time Emotion Recognition (HMR-TER) approach has been developed for enhancing e-learning, which was introduced by Du, Y. et al. [34]. By promptly offering feedback based on learners' facial expressions and vocal intonations, the e-learning was enhanced. This approach uses gesture recognition analysis to enhance participation and interaction and hybrid validation dynamic analysis to solve the low learner motivation. On the other hand, a few tests with a limited range of accuracy were made available. An Audio-Visual Automatic Speech Recognition (AV-ASR) system was proposed by S. Debnath et al. [35] to enhance the educational experience of those with physical disabilities by enabling hands-free computing. For visual speech data, the Local Binary Pattern-Three Orthogonal Planes (LBP-TOP) and Grey-Level Co-occurrence Matrix (GLCM) are suggested. The results of the study demonstrate that the suggested system obtains 96.50% accuracy for audio speech recognition and 76.60% accuracy for visual speech. However, as cluster size grows, accuracy decreases. This is due to the fact that a larger cluster size exhibits a dispersed representation of the data, lowering accuracy.

An Automatic Speech Recognition (ASR) system for the Lithuanian language was reported by L. Pipiras et al. [36]. It depends on deep learning techniques and recognizes spoken words only based on their phoneme sequences. The ASR task is solved using two different encoder-decoder models: a conventional model and a model with an attention mechanism. These models' effectiveness has been evaluated in two tasks: extended phrase recognition and isolated voice recognition. With shorter input sequences, the proposed model works merely well; however, it struggles with larger sequences. Bidirectional Long Short-Term Memory (BiLSTM) neural network and Wavelet Scattering Transform with Support Vector Machine (WST-SVM) classifier were developed by A. Lauraitis et al. [37] to identify patients' speech impairments at the beginning of central nervous system disorders (CNSD). The voice recorder from the Neural Impairment Test Suite (NITS) has been employed to capture speech data. Pitch contours, auditory spectrograms, Mel-frequency cepstral coefficients (MFCC), Gammatone cepstral coefficients (GTCC), and Gabor (analytic Morlet) wavelets are the sources of features that are extracted. Although there is a significant association among phoneme and grapheme sequence lengths in the corpus of data, not all patterns have the same length.

Table I shows the summary of literature reviews based on speech emotion recognition.

A. Problem Definition

The collected literature works show that the overall outcome of the existing approaches and the recognition ability to classify speech got affected due to the inappropriate techniques for selecting the relevant features; further, the complexity faced by the model to evaluate the huge datasets and poor classification methods relies as the reason to diminish the classification accuracy. Moreover, the results achieved superior values in training stage, while got affected in testing stage to classify the corresponding voice signal. Therefore, this research focused on two stages (i.e., feature selection and classification) which probably improvises the SER ability and helps to achieve better results in terms of accuracy, precision, recall and F1-score.

III. SER Using CP-LSTM With Attention Mechanism

The SER is a well-versed area which helps to perform communication among computers and humans. However, due to certain factors, overall efficiency of the recognition systems get diminished with poor accuracy. Therefore, more efforts have been put forward to enhance the efficiency during speech recognition but it had faced issues due to presence of undesired noises in the input signal. So, this research developed an effective framework using optimization-based feature selection and a novel deep learning approach. Initially, the data is

TABLE I. Summary of Literature Review

Author	Methodology	Disadvantage
Yildirim et al. [21]	Modified feature selection approach for speech emotion recognition.	Performing an exhaustive search for all subset space is infeasible.
Agarwal and Om [22]	An optimized Deep Neural Network (DNN) for speech emotion recognition.	Implementing a greater number of optimization techniques results in computational complexity.
Manohar and Logashanmugam [23]	A hybrid Deer Hunting with Adaptive Search (DH-AS) along with feature selection for recognizing the speech emotion.	The suggested approach was suitable only for the balanced dataset and complex due to its recurrent nature.
Mustaqeem and Kwon [24]	An optimal feature selection approach using two stream Deep Convolutional Neural Network (D-CNN).	The improper pre-processing leads to discrepancies and redundancies that diminishes the efficiency of the overall model.
Chattopadhyay et al. [25]	A hybrid feature selection approach using Clustering based Equilibrium Optimizer and Atom Search Optimization (CEOAS) for recognizing emotions from speech signals.	CEOAS exhibits premature convergence at the time of iterative searching process.
Kanwal et al. [26]	A Density based Spatial Clustering with Noise Genetic Algorithm (DGA) to recognize the type of emotion from the speech.	The PCA was computationally imprecise for huge datasets.
Barsainyan and Singh [27]	An optimized speech emotion recognition using 1D CNN.	The suggested framework was vulnerable to higher datasets due to limited model training.
Sun et al. [28]	A speech emotion recognition approach using Improved Masking based Empirical Model Decomposition and Convolutional Recurrent Neural Network (IMEMD-CRNN).	The mode mixing occurs which diminishes the capability of IMEMD to decompose the signal.
Ottoni et al. [29]	A deep learning approach based on CNN and Long Short-Term Memory (LSTM) which is used to recognize the speech emotions based on meta learning approach.	The lack of feature selection leads to computational complexity and diminishes the efficiency of classification.
Nhat Truong Pham et al. [30]	A Hybrid Data Augmentation (HDA) with Modified Attention-BasedDilated Convolutional and Recurrent Neural Network (mADCRNN) methods are proposed for speech emotion recognition.	The number of layers and parameters in the suggested mADCRNN model made it complicated.
Zengzhao Chen et al. [31]	A parallel network for multi-scale SER based on connection attention mechanism (AMSNet) is proposed for speech emotion recognition.	Because of their less conspicuous characteristics, neutral emotions are less often recognized, since the characteristics of neutral emotions were not sufficiently evident and it was challenging to categorize the corresponding voice signal.
Mustaqeem Khan et al. [32]	Multimodal Speech Emotion Recognition (MSER).	In order to address the restriction of the complicated interaction between speech signals and transcripts, these investigations disregarded the limitations of past knowledge.
Dal Ri et al. [33]	An extensive validation using CNN for speech emotion recognition.	An appropriate selection of features was mandatory for further enhancing the recognition.
Y. Du et al. [34]	Heuristic Multimodal Real-Time Emotion Recognition (HMR-TER) approach.	A few tests with a limited range of accuracy were made available.
S. Debnath et al. [35]	An Audio-Visual Automatic Speech Recognition (AV-ASR) system was proposed to enhance the educational experience of those with physical disabilities by enabling hands-free computing.	As cluster size grows, accuracy decreases. This is due to the fact that a larger cluster size exhibits a dispersed representation of the data, lowering accuracy.
L. Pipiras et al. [36]	An Automatic Speech Recognition (ASR) system was developed for the Lithuanian language.	With shorter input sequences, the proposed model works merely well; however, it struggles with larger sequences.
A. Lauraitis et al. [37]	Bidirectional Long Short-Term Memory (BiLSTM) neural network and Wavelet Scattering Transform with Support Vector Machine (WST-SVM) classifier.	Although there is a significant association among phoneme and grapheme sequence lengths in the corpus of data, not all patterns have the same length.

acquisitioned from five different datasets and it is pre-processed to neglect the undesired information from the audio signal. After this, the pre-processed data is used to acquire the features. The extracted features are provided to phase of selecting features that takes place using Improved Jellyfish Optimization Algorithm (IJOA). As the final stage, classification takes place using the Convolutional Peephole Long Short-Term Memory (CP-LSTM) with attention mechanism. The overall process involved in the proposed Speech Emotion Recognition (SER) framework is presented in Fig. 1.

A. Dataset Description

The SER proposed in this research utilized five different types of datasets such as Berlin Database of Emotional Speech (EMO-DB) [38], Ryerson Audio Visual Database of Emotional Speech and Song (RAVDESS) [39], Interactive Emotional Dyadic Motion Capture (IEMOCAP) [40], Surrey Audio-Visual Expressed Emotion (SAVEE) [41] and Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [42]. The aforementioned datasets are described as follows:

EMO-DB: It is obtained from Berlin emotion speech corpus that is recorded by a total of 10 actors where five were males and the remaining five were females. The dataset is comprised with a total of 535 audio files with an average time of 3-5 seconds and a sampling rate of 16kHz. Moreover, it is one of the vastly used datasets used in machine learning and deep learning techniques due to its clarity, which facilitates high recognition rates.

RAVDESS: It is one of the newly launched databases that is vastly utilized in evaluating the emotion of speech. This dataset is comprised

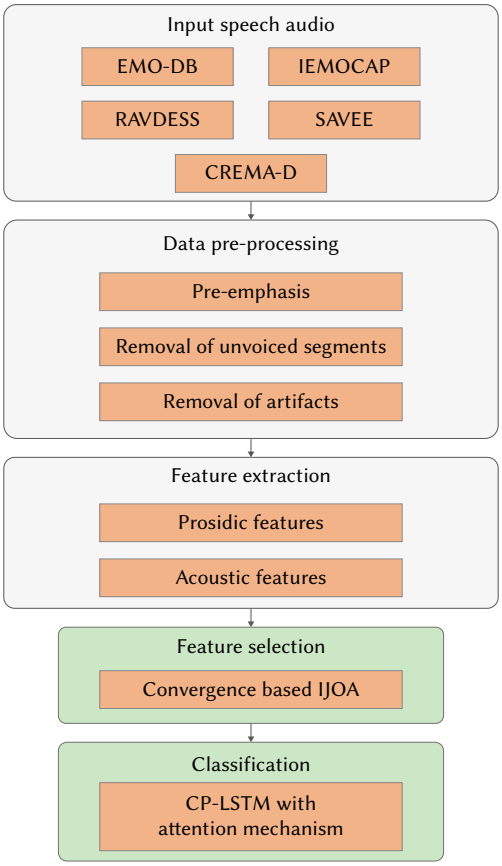

Fig. 1. Workflow of the proposed SER framework.

with total of eight kind of emotions which are recorded by 24 actors in a total of 12 sessions. The sampling rate of RAVDESS dataset is 48kHz with an average time of 3.5 s.

IEMOCAP: It is another type of vastly utilized SER datasets, which consists of improvised and scripted dialogues obtained from 10 actors in 5 sessions. This dataset is comprised with the audio clip of 12 hours and the utterances are annotated with emotion labels. IEMOCAP dataset is comprised with four stages such as anger, sad, neutral and happiness.

SAVEE: This dataset is comprised with a total of 480 utterances with varying emotions recorded by 4 actors at Centre for Vision, Speech and Signal Processing (CVSSP). Every speaker speaks 120 phonetically balanced English sentences based on 7 emotional classes.

CREMA-D: This dataset comprises 7,442 original clips from 91 actors. These clips have been gathered from 48 male and 43 female actors among the ages of 20 and 74 from a diversity of races and ethnicities (African America, Asian, Caucasian, Hispanic, and Unspecified).

B. Data Pre-Processing

Next to data acquisition, the pre-processing is performed to neglect the undesired information from the raw data. In this research, the data pre-processing is performed using pre-emphasis, removal of artifacts and removal of unvoiced segments. The process involved in the aforementioned techniques are presented in the following sub-sections.

1. Pre-Emphasis

The speech signal pre-emphasis is the initial level of pre-processing in higher frequency. Here, the speech signal is transferred via high pass filter for enhancing the high-frequency band's amplitude. In specific, this research makes use of Finite Response (FIR) filter that helps in the process of flattening high frequency audio signal. The process involved in pre-emphasis is represented in (1).

$$P(y) = 1 - hy^{-1} \tag{1}$$

Where the pre-emphasized output is represented as $P(y)$, the audio signal is represented as y and the filter co-efficient is represented as h.

2. Removal of Artifacts

After flattening the high frequency signals, the process of removing artifacts takes place which is used to remove the useless data and helps in effective recognition of emotions. In this research, the artifact removal takes place using the Fast Fourier Transform (FFT) that helps to remove the useless artifacts so the power spectrum of individual signals is evaluated using (2) as follows:

$$T_m(e) = \frac{1}{V}|S_m(e)|^2 \tag{2}$$

Where the spectral power of improved speech signals are represented as $S_m(e)$ and the total number of audio samples is represented as V.

3. Removal of Unvoiced Segments

Next to the stage of artifact removal, the unvoiced segments in the audio signals are removed, which increases the computational complexity while processing the audio signals. This research utilized Zero Cross Rate (ZCR) to remove the unvoiced segments. ZCR offers transition of signal over zero line which denotes noiseless measure in speech signals. ZCR is evaluated using (3) as follows:

$$Z_t = \frac{1}{2}\sum_{n=1}^{N}|sign(x[n]) - sign(x[n-1])| \tag{3}$$

Where *sign* is the function that presents 1 for positive arguments and 0 for negative arguments. The time domain signal at the frame t is denoted as $x[n]$ which provides a measure of noiseless signal.

Thus, the pre-processing performed based on the fore mentioned techniques helps to remove the undesired information from the audio signal and helps to achieve better classification results thereby minimizing the complexity of the data. The pre-processed data is provided to the phase of extracting features.

C. Feature Extraction

Next to pre-processing, the feature extraction is performed to extract the relevant features from the pre-processed output. This extraction of features will be based on prosodic and acoustic features. The prosodic features such as energy, entropy, pitch and formants are utilized to extract the relevant features. In a similar way, the acoustic features such as Linear Predictive Coding (LPC), Linear Prediction Cepstral Co-efficient (LPCC), Mel-Frequency Cepstral co-efficient (MFCC), spectral flux and Zero Cross Rate (ZCR) are extracted from the pre-processed output.

1. Prosodic Features

The prosodic features such as energy, entropy and pitch are considered. Among the three, energy acts as the fundamental speech signal processing. In emotion recognition, energy plays a major role in recognizing the speech signals. Secondly, pitch acts as the periodic standard where high frequency harmonics are captured. The characteristics can be retained for every individual frame and it is pre-processed using short term analysis approach. The energy (E) , pitch (P) and entropy (H) are evaluated based on (4)-(6) respectively.

$$E = \frac{1}{N}\sum_{n=m-N_w+1}^{m}|f_s(n;m)|^2 \tag{4}$$

$$P = \frac{F}{N}\text{argmax}_{\eta}\{|r(\eta,m)|\}_{\eta=N_w(F_l/F_s)}^{\eta=N_w(F_h/F_s)} \tag{5}$$

$$H = -\sum_{i=1}^{N} s(i) \times \log[s(i)] \tag{6}$$

Where energy of speech frame is denoted as E, total count of frames is denoted as N and number of samples in the frame is denoted as N_w. The entropy is denoted as H, pitch frame is represented as P and the sampling frequency is represented as F_s. The low pitch frequency and high pitch frequency is represented as F_l and F_h respectively, measurement metric of glottal velocity is denoted as argmax.

2. Acoustic Features

Other than prosodic features, the acoustic features are considered while extracting the features. The fore mentioned acoustic features such as LPC, LPCC, MFCC, spectral flux, and ZCR are considered while extracting the acoustic based features.

LPC: The LPC computes the intensity of the input signal thereby eliminating the frequency of left over buzz. The remaining portion of the signal is achieved as residual signal that is represented in (7):

$$s(n) = \alpha_1 s(n-1) + \alpha_2 s(n-2) + \cdots + \alpha_p s(n-p) \tag{7}$$

Where the linear co-efficient is denoted as $\alpha_1, \alpha_1, \ldots \alpha_p$.

LPCC: It is similar to the feature extraction performed by LPC where the Cepstral co-efficient is extracted from the features of LPC. Next to this, the representation of co-efficient is performed using the derivative of Fourier transform.

MFCC: It is the illustration of short time power spectrum of sound where the square magnitude of windowed speech signal is evaluated. The Mel scale filters along with log power spectrum is employed by overlapping the critical band filters. The co-efficient is assessed by performing Discrete Cosine Transform (DCT) of Mel-bin log energies is denoted in equation (8).

$$C_i(t) = \sum_{a=1}^{B} \log m_a(t) . \cos\left(\frac{i(a-0.5)\pi}{b}\right) \tag{8}$$

Where, Mel cepstral coefficients are denoted as C_i, an amount of Mel filters at the filter bank is specified as B, $i = 1, 2, \ldots, p$, amount of DCT points is denoted as p and the triangular filter bank function is represented as b.

Spectral flux: The speech signals which are pre-processed are obtained using the spectral flux to extract the various features of emotions. The spectral flux of the speech signal is denoted as SF_m which is evaluated using (9) as follows:

$$SF_m = \sum_{G=0}^{\frac{pt}{2}-1} (|E_m(G)|^2 - |E_{m-1}(G)|^2) \tag{9}$$

Where the spectrum value of the speech signal at frame m and $m-1$ in frequency bin G are represented as $E_m(G)$ and $E_{m-1}(G)$ respectively, and total amount of points in the spectrum is denoted as pt.

ZCR: It is a general type of feature insert which quantifies the amplitude of the speech signal, which has a zero value threshold in a particular time frame. ZCR has the ability to distinguish among the voiced and unvoiced signals and mathematically ZCR is evaluated based on (10) as follows:

$$ZCR = \frac{1}{T-1} \sum_{t=1}^{T-1} 1_{R<0} \tag{10}$$

Where the length of signal S is represented as T and $1_{R<0}$ is the indicator function.

Finally, the features which are extracted based on the prosodic features and the acoustic features are concatenated which is fed into the stage of feature selection to select the optimal features.

D. Feature Selection

Next to the stage of extraction of features, the selection of optimal features is performed with the help of proposed Improved Jellyfish Optimization Algorithm (IJOA). The existing Jellyfish Optimization Algorithm (JOA) is based on the searching pattern while searching for the food. Initially, the jelly fish goes through ocean current then moves within the group based on two motions specified as type A and type B. A brief explanation about the iterative process involved in IJOA along with the enhancement made to select the relevant features are represented in the following sub-sections.

1. Overview of JOA

The movement of jelly fish is based on their active and passive behavior represented as type A motion and type B motion. The time control principle plays a major role in determining those two time varying motion types and this plays an important role in the development of controlling variation among type A and type B.

a) Initialization of Population

In general, the population of optimization algorithms are initialized in a random manner. This randomized initialization results in minimal precision and limited running value. So in JOA, the initialization is performed using logistic maps and randomness is generated using chaotic maps. This is represented in (11) as follows:

$$P_{i+1} = \eta P_i (1 - P_i) \tag{11}$$

Where logistic chaos value based on candidate's position is represented as P_i and the population at initial stage is represented as P_0.

b) Behavior of Following Ocean Current

The direction of each variable for candidate solution from the position to optimal position is represented as current direction ($\overrightarrow{Direction}$) which is expressed in (12) as follows:

$$\overrightarrow{Direction} = \frac{1}{N} \sum \overrightarrow{Direction_i} = \frac{1}{N} \sum (P^* - e_c P_i) = P^* - e_c \frac{\sum P_i}{N} = P^* - e_c \mu = P^* - df \tag{12}$$

Where N, e_c and μ represents population of individual candidates, attraction factor and position of jelly fish correspondingly. The optimal position of individual candidate is represented as P^* and the variation among the optimal and the average location is represented as $df = e_c \mu$.

c) Movement of Jelly Fish

The movement of jelly fish is based on two types such as type A and type B. In type A, most of the candidate solutions do not show ability and in type B, the jelly fish starts the move in passage of time.

(1) Movement of type A

It is the type of passive movement where individual candidate changes along the own position which is updated using (13) as follows:

$$P_i(t+1) = P_i(t) + \gamma \times r_3 \times (U_b - L_b) \tag{13}$$

Where upper and lower limit of the search space is denoted as U_h and L_b correspondingly, movement factor is represented as γ and the value of r_3 lies among the range of (0,1).

(2) Movement of type B

It is the type of active movement where the individual candidate (j) is selected in a random manner. When the total quantity of food exceeds the location of selected candidates, the position of P_j exceeds the own location P_i. Every individual candidate migrates from one direction to another in search of food where the position of the candidate gets updated on the basis of (14)

$$P_i(t+1) = P_i(t) + \overrightarrow{step} \tag{14}$$

Where $\overrightarrow{step}$ is calculated as expressed in equations (15) and (16).

$$\overrightarrow{step} = rand\ (0,1) \times \overrightarrow{DDirection} \tag{15}$$

$$\overrightarrow{DDirection} = \begin{cases} P_j(t) - P_i(t)\ if\ f(P_i) \geq f(P_j) \\ P_i(t) - P_j(t)\ if\ f(P_i) < f(P_j) \end{cases} \tag{16}$$

d) Time Control Mechanism

After the stage of capturing the movement, time control theory is utilized to adjust the varying candidate solutions. The candidate individuals present in the population of ocean current are denoted in (17).

$$C(t) = |(1 - t/T) \times (2 \times rand(0,1) - 1)| \quad (17)$$

Where the iterations at current stage and the maximum stage is denoted as t and T correspondingly.

2. Convergence Based IJOA

The existing JOA faced issues related to poor precision value and tends to fall into the local optima. So, this research introduced improvisation in JOA by introducing sine and cosine learning factors, and a premature convergence strategy. The sine and cosine learning is based on stage of exploration and the premature convergence strategy is based on exploitation stage. The inducement of learning strategy and the premature convergence strategy helps to overcome the issues related to poor precision and poor convergence rate.

a) Factors Based on Sine and Cosine Learning for Exploration

In the exploration phase, jelly fish performs type B motion in the whole population. The jelly fish learns from the other individual in current population that results in certain inappropriateness and lacks a proper exchange among the population. This results in placement of proper candidate solution and slows down the convergence speed. Therefore, the learning factors such as ω_1 and ω_2 based on sine and cosine functions are introduced to enhance the capability of jelly fish to learn from random and best individuals. The presentation of learning strategies in the exploration stage helps to enhance the quality of candidate solution by identifying the optimal location and helps to improvise the convergence speed. The mathematical equations of sine and cosine learning factor are represented as ω_1 and ω_2 which are presented in (18) and (19) as follows:

$$\omega_1 = 2.\sin\left[\left(1 - \frac{t}{T}\right).\pi/2\right] \quad (18)$$

$$\omega_2 = 2.\cos\left[\left(1 - \frac{t}{T}\right).\pi/2\right] \quad (19)$$

During the stage of updating the type B movement, the location of the jelly fish got updated as is denoted in (20)

$$P_i(t+1) = \omega_1.\left(P_i(t) + \overrightarrow{step}\right) + \omega_2.\left(P^* - P_i(t)\right) \quad (20)$$

Where the value of $\overrightarrow{step}$ is represented in (15) and (16). The actual JOA utilizes a random learning technique to learn from their individual. Moreover, the poor fitness functions of the jelly fish limit the speed of convergence. Then, the sine and cosine factors help to learn from random solutions and help to enhance the quality of solution with quick convergence rate.

b) Premature Convergence Strategy for Exploitation

The existing JOA experience low exploitation due to the inability of algorithm to accelerate the convergence for an optimal solution. Additionally, the capacity of swarm exploitation for accomplishing a search in the population is restricted. This problem occurs in exploitation, that is important in updating solutions and enhancing the search in the population. The capacity of the algorithm is restricted, when the small solution is enhanced, and the capacity of swarm for searching other area maximizes when the r is high. The mathematical expression for premature convergence rate is denoted in (21).

$$\overrightarrow{X_i}(t+1) = \overrightarrow{X_i}(t) + r \times \left(\vec{X}_{r1}(t) - \vec{X}_{r2}(t)\right) + (1-r) \times \left(X^* - \vec{X}_{r3}(t)\right) \quad (21)$$

Where the indices of three solutions which are picked up in a random manner from $r1$, $r2$ and $r3$, where the control parameter is represented as r. The value of control parameters lies in the range of 0 and 1 which is utilized in controlling the movement of current solution. The premature convergence strategy is used to accelerate the rate of convergence with randomly selected solutions from the population. This entire process involved in selecting the optimal features is performed using the proposed IJOA and emotion categorization from speech is carried out using CP-LSTM with attention mechanism.

E. Classification Using CP-LSTM With Attention Mechanism

The various forms of Long Short-Term Memory (LSTM) with various architectures are vastly utilized in the applications related to speech emotion recognition. The traditional architecture of LSTM is widely utilized in SER. However, there are issues because dependencies among the cells are not strong, which affects the overall classification efficiency. Then, this research introduced an effective classification approach using the proposed CP-LSTM with attention mechanism. Even with the closed output gate, the suggested CP-LSTM permits access to the previous cell state. Moreover, the content of previous memory cell helps to capture the complete dependency to improve the model accuracy. Also, the attention strategy is included in the final layer of CP-LSTM model that assists to choose the important term and capturing the complete pattern of training data. The addition of attention layer supports the model to generate syntactically and semantically [43]; the brief details about the traditional architecture of LSTM and the proposed CP-LSTM with attention mechanism for SER is explained in following sections.

1. Convolutional Peephole LSTM With Attention Mechanism

The LSTM is a kind of Recurrent Neural Network (RNN) which has the capability to hold and remember the information for a particular period of time. The LSTM is highly recommended in processing the sequences from the selected features. The architectural diagram of the LSTM model is presented in Fig. 2.

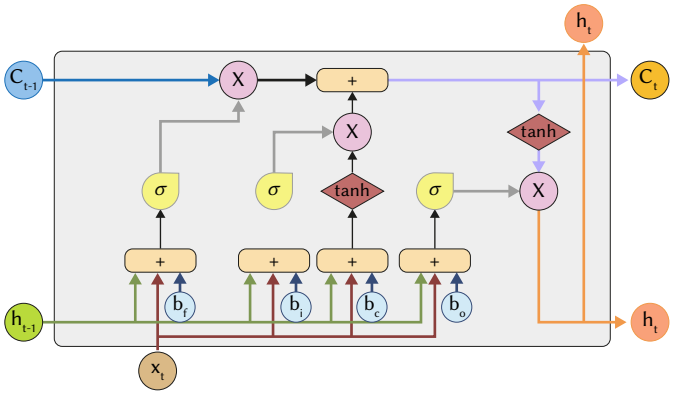

Fig. 2. Architectural diagram of the traditional LSTM model.

The architecture of LSTM is comprised with memory cells and three gates like forget, input and output gates in which data is stored using the memory cell and the control cell states are controlled using the remaining three gates. From Fig. 1, the line which is passed over architecture is represented as a memory pipe. The memory in previous state of memory line is denoted as C_{t-1} and the memory pipe is organized based on three gates. The sigmoid function is used to convert output of forget gate from 0 to 1. The bitwise summation integrates the temporary memory which is created using input gate with prior memory state along with the final memory which is represented as C_t.

The forget gate is a single layered neural network which takes place in different operation and regulate the memory content of previous cell. The forget gate obtains output state h_{t-1}, input vector x_t and bias input b_f. The final output generated from the forget gate is represented as f_t among the range of 0 to 1 and it is multiplied with C_{t-1}. The lower value of f_t prohibits the content of previous memory where the high value of f_t permit prior memory state to contribute the current state. The forget gate is represented in (22).

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \tag{22}$$

The temporary memory state is comprised with sigmoid function and hyperbolic tangent function where sigmoid function produces output i_t, which is integrated in previous state, and the *tanh* is a kind of activation function whose value lies among 0 and 1. The temporary memory with large *tanh* function contributes better to the memory cell which is represented in (23) and (24).

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \tag{23}$$

$$C_t = f_t C_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{24}$$

The output gate distinguishes the content h_t at a time t and the value of o_t in the range 0 to 1. When the value of O_t is equal to 1 then $h_t = C_t$ where the entire C_t is passed to next state h_t. The outcome through the output gate and hidden state of succeeding layer is represented as o_t and h_t which is represented in (25) and (26) respectively.

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \tag{25}$$

$$h_t = o_t \tanh(c_t) \tag{26}$$

Based on the architecture of LSTM, the control gates are not incorporated with memory cell. Moreover, the output gate of LSTM is in closed state during training time, which may result in inappropriate classification result in SER. Therefore, this research introduced an advanced LSTM architecture known as CP-LSTM with attention mechanism to rectify the issues aroused while performing emotion recognition from speech.

a) CP-LSTM With Attention Mechanism for SER

The traditional LSTM architecture faced issues related to poor configuration among gates which prohibits the memory usage of prior memory states. The fore mentioned issue in traditional LSTM can be overwhelmed by introducing connection among individual gate and memory content which is referred as CP-LSTM. The peephole connection of CP-LSTM helps all the gates to access memory content. In CP-LSTM, the previous cell state C_{t-1} is linked with the controlling gate which is referred as peephole connections and the presence of this peephole permits an additional parameter and a memory state as the input of CP-LSTM. The inclusion of this additional input in each gate allows admission to memory content of prior cell state. The architectural diagram of CP-LSTM with attention mechanism is represented in Fig. 3. The architecture of CP-LSTM is similar to the architecture of traditional LSTM which is based on the mathematical expressions listed in (27)-(31)

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{27}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \tag{28}$$

$$C_t = f_t C_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{29}$$

$$O_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{30}$$

$$h_t = o_t \tanh(C_t) \tag{31}$$

The proposed CP-LSTM utilized memory content of prior cell as input and the connection of C_{t-1} in CP-LSTM enhances the accuracy of prediction tasks. The attention mechanism present in the CP-LSTM architecture evaluates weight for every individual word which is based on probability function that is used to evaluate the significant factors of input. The attention value of the signal is based on how much attention need to be paid while generating the recognition output from speech signals. When the speech signal is passed to the final hidden layer, the average weight of the speech signal is evaluated. Next to this, it is passed to the softmax layer with memory content of final hidden layer to predict emotion from speech. The attention weight of input features of speech signal is evaluated based on (32) as follows:

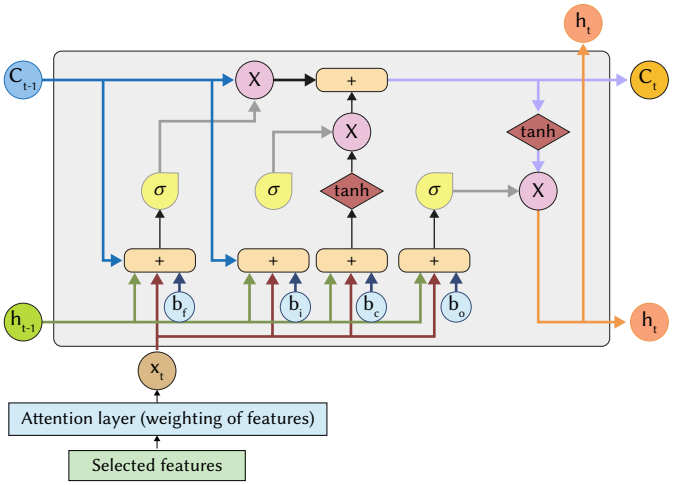

Fig. 3. Architectural diagram of CP-LSTM with attention layer.

$$a_{y_t'}(t) = \frac{\exp\left(h_{x_t}^T h_{y_{t'}}\right)}{\Sigma_t^T \exp\left(h_{x_t}^T h_{y_{t'}}\right)} \tag{32}$$

Where last hidden layer which is created next to processing features is represented as $\exp\left(h_{x_t}^T\right)$. The attention mechanism in the CP-LSTM architecture helps to handle the problems of mapping large source to a static length. Moreover, the softmax layer is utilized to remove the outliers from output and map the vector. Thus, combination of proposed CP-LSTM with attention mechanism helps in an effective SER by recognizing the signal pattern, which results in better classification accuracy.

IV. Results and Analysis

This section presents the analysis of the results achieved by evaluating the proposed CP-LSTM with attention mechanism with the state of art techniques and the existing approaches. Moreover, the efficiency of IJOA is evaluated with state of art optimization techniques while selecting the relevant features.

A. Experimental Setup and Evaluation Metrics

The efficiency of the CP-LSTM with attention mechanism based on the features selected using IJOA is implemented in python software and the system has configuration such as 16 GB RAM, i7 processor and windows 11 OS. The efficiency of the proposed approach is evaluated by considering the performance metrics such as accuracy, precision, recall and F1 score. Table II presents the performance metrics and the respective formula used while evaluation.

TABLE II. Evaluation Metrics

Metrics	Formulae
Accuracy (A)	$A = {TP + TN}/{TP + TN + FP + FN}$
Precision (P)	$P = {TP}/{TP + FP}$
Recall (R)	$R = {TP}/{TP + FN}$
F-1 score (F)	$F = 2 \times {P \times R}/{P + R}$

Where *TP* and *TN* denote true positives and true negatives whereas the false positives and false negatives are represented as *FP* and *FN* respectively.

B. Performance Analysis

In this section, the performance of the proposed CP-LSTM with attention mechanism along with the proposed feature selection approach using IJOA is evaluated. The datasets EMO-DB, IEMOCAP, RAVDESS and SAVEE are used to evaluate the efficiency of IJOA and CP-LSTM. The quantitative analysis of the proposed method with all five datasets is presented in Table III.

TABLE III. Quantitative Evaluation of Proposed Approach

Datasets	Accuracy (%)	Precision (%)	Recall (%)	F-1 score (%)
EMO-DB	99.59	98.76	98.21	99.82
IEMOCAP	99.88	98.12	98.40	98.36
RAVDESS	99.54	99.83	99.57	98.36
SAVEE	98.89	98.57	98.19	98.43
CREMA-D	99.12	98.83	98.76	98.81

The Table III shows that the proposed approach accomplished an accuracy of 99.59%, 99.88%, 99.54%, 98.89% and 99.12% respectively for EMO-DB, IEMOCAP, RAVDESS, SAVEE and CREMA-D datasets. The following section presents the detailed analysis of the experimental outcome possessed by IJOA and CP-LSTM with attention mechanism for the different datasets.

1. Evaluation Based on Feature Selection

In this sub-section, the performance of IJOA utilized in feature selection is evaluated with the state of art optimization techniques such as Whale Optimization Algorithm (WOA), Grasshopper Optimization Algorithm (GOA) and Jellyfish Optimization Algorithm (JOA). The evaluation is performed by considering the five datasets utilized in this research. First, the efficiency of IJOA is evaluated with the existing optimization techniques such as WOA, GOA and JOA for EMO-DB dataset. The experimental outcome achieved while evaluating IJOA with existing ones for EMO-DB dataset is presented in table IV.

TABLE IV. Evaluation of IJOA for EMO-DB Dataset

Method	Accuracy (%)	Precision (%)	Recall (%)	F-1 score (%)
WOA	97.35	96.22	96.81	96.05
GOA	97.62	96.75	96.87	96.27
JOA	96.81	97.20	96.15	96.69
IJOA	99.24	97.61	97.48	96.81

Experimental outcome from Table IV shows that IJOA achieved better results than the state of art optimization techniques. The accuracy of the proposed IJOA is 99.24% which is comparably higher than WOA, GOA and JOA with accuracies of 97.35%, 97.62% and 96.81%. Fig. 4 shows the graphical depiction of IJOA's performance for EMO-DB dataset.

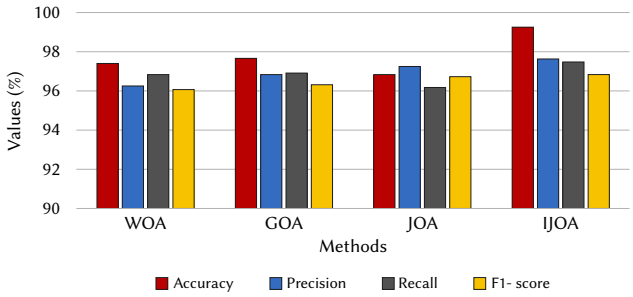

Fig. 4. Graphical representation of optimization techniques performance on EMO-DB dataset.

Secondly, the performance of IJOA is evaluated with WOA, GOA and JOA for IEMOCAP dataset. Table V shows the experimental results achieved while evaluating IJOA with state of art optimization techniques.

TABLE V. Evaluation of IJOA for IEMOCAP Dataset

Method	Accuracy (%)	Precision (%)	Recall (%)	F-1 score (%)
WOA	96.12	95.23	95.39	95.28
GOA	95.66	96.21	96.58	96.59
JOA	97.18	96.68	96.19	96.17
IJOA	99.37	99.82	99.76	99.10

Table V shows IJOA achieved better results than state of art optimization techniques. The accuracy of the proposed IJOA is 99.37% which is comparably higher than WOA, GOA, and JOA with accuracies of 96.12%, 95.66% and 97.18% respectively. Fig. 5 shows the graphical depiction of IJOA's performance for IEMOCAP dataset.

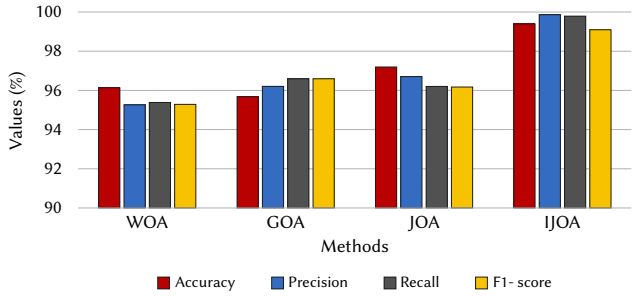

Fig. 5. Graphical representation of optimization techniques performance on IEMOCAP dataset.

Thirdly, the performance of IJOA is evaluated with WOA, GOA and JOA for RAVDESS dataset. Table VI shows the experimental results achieved while evaluating IJOA with state of art optimization techniques.

TABLE VI. Evaluation of IJOA for RAVDESS Dataset

Method	Accuracy (%)	Precision (%)	Recall (%)	F-1 score (%)
WOA	94.21	96.90	96.01	96.22
GOA	96.59	97.55	95.89	95.97
JOA	97.82	96.81	96.27	96.80
IJOA	99.98	98.81	99.89	99.82

Table VI exhibits IJOA achieved better results than state of art optimization techniques. The accuracy of the proposed IJOA is 99.98% which is comparably higher than WOA, GOA, and JOA with accuracies of 94.21%, 96.59% and 97.82% respectively. Fig. 6 shows the graphical depiction of IJOA's performance for RAVDESS dataset.

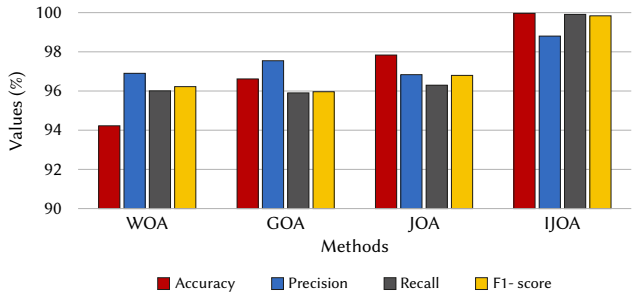

Fig. 6. Graphical representation of optimization techniques performance on RAVDESS dataset.

Finally, the performance of IJOA is evaluated with WOA, GOA and JOA for SAVEE dataset. Table VII shows the experimental outcome achieved while assessing IJOA with state of art optimization techniques.

TABLE VII. EVALUATION OF IJOA FOR SAVEE DATASET

Method	Accuracy (%)	Precision (%)	Recall (%)	F-1 score (%)
WOA	96.82	96.78	96.82	96.29
GOA	97.11	97.80	97.99	97.87
JOA	97.56	97.25	97.69	97.58
IJOA	98.81	98.66	98.85	99.82

Table VII exhibits IJOA achieved better results than state of art optimization techniques. The accuracy of the proposed IJOA is 98.81% which is comparably higher than WOA, GOA, and JOA with accuracies of 96.82%, 97.11% and 97.56% respectively. Fig. 7 shows the graphical depiction of IJOA's performance for SAVEE dataset.

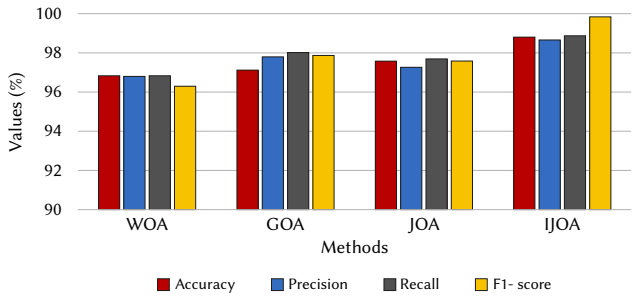

Fig. 7. Graphical representation of optimization techniques performance on SAVEE dataset.

The table VIII shows the experimental results accomplished while evaluating IJOA with state of art optimization methods on CREMA-D dataset.

TABLE VIII. EVALUATION OF IJOA FOR CREMA-D DATASET

Method	Accuracy (%)	Precision (%)	Recall (%)	F-1 score (%)
WOA	95.38	94.18	95.44	95.26
GOA	95.91	95.73	96.19	96.43
JOA	96.56	96.49	96.38	96.71
IJOA	98.91	98.19	97.91	98.05

Table VIII shows that IJOA attained best results than state of art optimization methods. The accuracy of the proposed IJOA is 99.91% which is greater than WOA, GOA, and JOA with accuracies of 95.38%, 95.91% and 96.56%, respectively. Fig. 8 displays the graphical depiction of IJOA's performance for CREMA-D dataset.

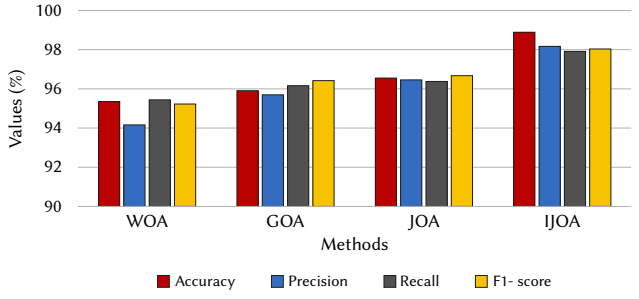

Fig. 8. Graphical representation of IJOA's performance on CREMA-D dataset.

The overall results from Table IV – VIII show that the proposed IJOA achieved better results in overall performance metrics when it is compared with existing optimization techniques. For example, the accuracy of the proposed IJOA for SAVEE dataset is 98.81% which is comparably higher than WOA, GOA, and JOA methods. The optimal outcome of proposed IJOA is due to inclusion of the sine and cosine learning factors for exploration and the premature convergence strategy for stage of exploitation.

2. Evaluation Based on Classifier

In this sub-section, the performance of the proposed CP-LSTM with attention mechanism is evaluated based on its efficiency in SER with the five datasets utilized in this research. The performance of CP-LSTM with attention mechanism is evaluated with Recurrent Neural Network (RNN), Deep Belief Network (DBN) and Long Short Term Memory (LSTM) with attention mechanism. The performance is evaluated with the five publicly available datasets EMO-DB, IEMOCAP, RAVDESS, SAVEE and CREMA-D. Table IX shows the results achieved by the suggested classifier for EMO-DB dataset.

TABLE IX. COMPARISON OF DIFFERENT CLASSIFIERS WITH ATTENTION MECHANISM FOR EMO-DB DATASET

Method	Accuracy (%)	Precision (%)	Recall (%)	F-1 score (%)
RNN	96.11	95.23	95.90	95.57
DBN	95.26	96.20	96.96	96.58
LSTM	97.02	96.53	96.62	96.57
CP-LSTM	99.59	98.76	98.21	99.82

The classification accuracy of CP-LSTM with attention mechanism is 99.59% which is comparably higher than the existing RNN, DBN and LSTM with accuracies of 96.11%, 95.26% and 97.02%.

Secondly, the performance of CP-LSTM with attention mechanism is evaluated for IEMOCAP dataset. Table X shows the results achieved by the suggested classifier for IEMOCAP dataset.

TABLE X. COMPARISON OF DIFFERENT CLASSIFIERS WITH ATTENTION MECHANISM FOR IEMOCAP DATASET

Method	Accuracy (%)	Precision (%)	Recall (%)	F-1 score (%)
RNN	96.35	97.91	96.12	96.61
DBN	96.82	98.53	97.34	96.62
LSTM	97.25	97.75	97.96	97.42
CP-LSTM	99.88	98.12	98.40	98.36

The classification accuracy of CP-LSTM with attention mechanism is 99.88% which is comparably higher than the existing RNN, DBN and LSTM with accuracies of 98.12%, 98.12% and 98.40%.

Thirdly, the performance of CP-LSTM with attention mechanism is evaluated for RAVDESS dataset. Table XI shows the results achieved by the suggested classifier for RAVDESS dataset.

TABLE XI. COMPARISON OF DIFFERENT CLASSIFIERS WITH ATTENTION MECHANISM FOR RAVDESS DATASET

Method	Accuracy (%)	Precision (%)	Recall (%)	F-1 score (%)
RNN	97.23	97.87	98.33	97.59
DBN	98.28	95.64	95.80	95.87
LSTM	97.53	98.82	98.97	98.48
CP-LSTM	99.54	99.83	99.57	98.36

The classification accuracy of CP-LSTM with attention mechanism is 99.54% which is comparably higher than the existing RNN, DBN and LSTM with accuracies of 97.23%, 98.28% and 97.53%. Table XII shows the results achieved by the suggested classifier for SAVEE dataset.

TABLE XII. Comparison of Different Classifiers With Attention Mechanism for SAVEE Dataset

Method	Accuracy (%)	Precision (%)	Recall (%)	F-1 score (%)
RNN	96.21	95.09	97.56	96.41
DBN	96.09	97.02	96.78	96.89
LSTM	97.78	97.54	96.99	97.57
CP-LSTM	98.89	98.57	98.19	98.43

The classification accuracy of CP-LSTM with attention mechanism is 98.89% which is comparably higher than the existing RNN, DBN and LSTM with accuracies of 96.21%, 96.09% and 97.78%. Finally, the performance of CP-LSTM with attention mechanism is evaluated for CREMA-D dataset which is shown in Table XIII.

TABLE XIII. Comparison of Different Classifiers With Attention Mechanism for CREMA-D Dataset

Method	Accuracy (%)	Precision (%)	Recall (%)	F-1 score (%)
RNN	95.83	95.16	96.16	95.93
DBN	96.76	96.43	96.82	96.59
LSTM	97.98	97.59	97.38	97.17
CP-LSTM	99.12	98.83	98.76	98.81

The classification accuracy of CP-LSTM with attention mechanism is 99.12% which is comparably higher than the existing RNN, DBN and LSTM with accuracies of 95.83%, 96.76% and 97.98%.

Table IX - Table XIII show the results achieved while evaluating the proposed classifier with the five publicly available dataset utilized in this research. The performance of the proposed classifier seems to be high when it is compared with the existing state of art classifiers. The outstanding result of the proposed classification approach is due the peephole connection of CP-LSTM helps all the gates to evaluate memory content. In CP-LSTM, the prior cell state linked with the controlling gate represented as peephole connections and the presence of this peephole permits an additional parameter and a memory state as the input of CP-LSTM. The attention mechanism in the CP-LSTM architecture helps to handle problems of mapping large fixed length. Thus, combination of the proposed CP-LSTM with attention mechanism helps in an effective SER by recognizing the signal pattern and results in better classification accuracy.

C. Comparative Analysis

In this section, the efficiency of the proposed classifier is assessed with existing techniques based on SER. The comparison is performed with existing techniques such as DNN-DHO [22], DH-AS [23], D-CNN [24], CEOAS [25], CNN-LSTM [29] and CNN [33]. Table XIV shows the results achieved while evaluating the suggested approach with existing techniques and different datasets.

In an overall, the suggested method attained better results than existing ones due to the effective feature selection performed using IJOA and the efficient classification performed using the proposed CP-LSTM with attention mechanism. For example, the classification accuracy of CP-LSTM for RAVDESS dataset is 99.54% which is comparably higher than DNN-DHO, DH-AS, D-CNN and CNN-LSTM with accuracies of 97.5%, 95.6%, 85% and 97.01% respectively.

D. Discussion

This sub-section provides a brief discussion about the results achieved while evaluating CP-LSTM along with its advantages. In performance analysis, the efficiency of CP-LSTM with attention mechanism is evaluated with five publicly available datasets (EMO-DB, IEMOCAP, RAVDESS, SAVEE and CREMA-D). The proposed IJOA used for feature selection outperforms state of art optimization techniques

TABLE XIV. Comparative Table

Method	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F-1 score (%)
DNN-DHO [22]	RAVDESS	97.5	97.52	97.75	97.19
DH-AS [23]	SAVEE	95.73	98.21	94.6	96.37
	RAVDESS	95.6	98.33	94.26	96.25
D-CNN [24]	EMO-DB	95	95.5	85	85.2
	SAVEE	82	74.8	72.9	73.7
	RAVDESS	85	85.4	85.4	85.4
CEOAS [25]	EMO-DB	98.72	DNA	DNA	DNA
	SAVEE	98.01	DNA	DNA	DNA
	IEMOCAP	74.25	DNA	DNA	DNA
CNN-LSTM [29]	SAVEE	90.62	DNA	DNA	DNA
	RAVDESS	97.01	DNA	DNA	DNA
CNN [33]	IEMOCAP	63	DNA	DNA	DNA
	RAVDESS	83	DNA	DNA	DNA
	CREMA-D	68	DNA	DNA	DNA
CP-LSTM with attention mechanism	EMO-DB	99.59	98.76	98.21	99.82
	SAVEE	98.89	98.57	98.19	98.43
	IEMOCAP	99.88	98.12	98.40	98.36
	RAVDESS	99.54	99.83	99.57	98.36
	CREMA-D	99.12	98.83	98.76	98.81

*DNA- data not available

such as WOA, GOA, and JOA. Similarly, CP-LSTM outperforms state of art classification approaches such as RNN, DBN and LSTM. The comparative analysis demonstrates that the CP-LSTM with attention mechanism offers better performance than the existing techniques such as DNN-DHO, DH-AS, D-CNN, CEOAS and CNN-LSTM. The sine and cosine functions incorporated in exploration stage of IJOA help to improve the candidate solution's quality while searching for optimum subset of features. This enhanced searching capacity of IJOA helps to mitigate the irrelevant features which is helpful in enhancing the recognition. On the contrary, the designed CP-LSTM allows the access to the previous cell state that leads to acquire the complete dependency for enhancing the SER performances. Further, the attention strategy included in the final layer of CP-LSTM is used for selecting the significant term and obtaining the complete pattern of training information for an additional improvement in the SER.

V. Conclusion

This research study introduced an effective classification approach using CP-LSTM with attention mechanism where the selected features are provided as input using IJOA. The major contribution of this research is to produce a robust framework for speech emotion recognition using five datasets namely, EMO-DB, IEMOCAP, RAVDESS, SAVEE and CREMA-D. These five datasets are pre-processed using pre-emphasis, removal of artifacts and removal of unvoiced segments to remove the undesired information from the signals. Then, the extraction of features takes place using prosodic features and acoustic features. The prosodic features such as energy, entropy, pitch and formants are utilized to extract the relevant features. In a similar way, the acoustic features such as LPC, LPCC, MFCC, spectral flux and ZCR are extracted from pre-processed output. Next to the feature extraction, the IJOA is used to select the relevant features which are fed into the classification that is performed using CP-LSTM with attention mechanism. The experimental results exhibit effectiveness of suggested approach by analyzing the classification accuracy for EMO-DB, IEMOCAP, RAVDESS and SAVEE datasets, which are 99.59%, 99.88%, 99.54% and 98.89% respectively, which is

comparably higher than the existing state-of-the-art models. Similarly, while working on RAVDESS dataset, the existing DNN-DHO, DH-AS, D-CNN and CNN-LSTM attained an accuracy of 97.5%, 95.6%, 85% and 97.01% respectively which is lower than proposed approach which has 99.54% accuracy. In the future, the suggested framework will be further implemented for time scenarios in industries and healthcare to analyze the classification accuracy.

Notation List

symbol	Description
$P(y)$	Pre-emphasized output
y	Audio signal
h	Filter co-efficient
$S_m(e)$	Spectral power of improved speech signals
V	Total number of audio samples
$sign$	Function
$x[n]$	Time domain signal
E	Energy of speech frame
P	Pitch frame
H	Entropy
N	Total count of frames
F_s	Sampling frequency
F_l	Low pitch frequency
F_h	High pitch frequency
argmax	Measurement metric of glottal velocity
$S(n)$	Residual signal
$\alpha_1, \alpha_1, \ldots \alpha_p$	Linear co-efficient
b	Triangular filter bank function
SF_m	Spectral flux of the speech signal
$E_m(G)$	Spectrum value of speech signal
S	Signal
T	Length of signal
$1_{R<0}$	Indicator function
P_i	Candidate's position
P_0	Population at initial stage is represented as
$(\overrightarrow{Direction})$	Direction of each variable
N	Population of individual candidates
e_c	Attraction factor
μ	Position of jelly fish
P^*	Optimal position of individual candidate
df	Variation among the optimal and the average location
U_b and L_b	Upper and lower limit
γ	Movement factor
j	Individual candidate
P_j	Position
P_i	Location
t	Current iterations
T	Maximum iteration
ω_1 and ω_2	Sine and cosine learning factor
r_1, r_2 and r_3	Random variables
r	Control parameter
C_{t-1}	Previous cell state
$\exp(h_{x_t}^T)$	Last hidden layer
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

Declarations

Funding: This work has been supported by the project PCI2022-134990-2 (MARTINI) of the CHISTERA IV Cofund 2021 program; by MCIN/AEI/10.13039/501100011033/ and European Union NextGenerationEU/PRTR for XAI-Disinfodemics (PLEC 2021-007681) grant, by European Comission under IBERIFIER Plus - Iberian Digital Media Observatory (DIGITAL-2023-DEPLOY- 04-EDMO-HUBS 101158511); and by EMIF managed by the Calouste Gulbenkian Foundation, in the project MuseAI.

Data Availability: The datasets generated during and/or analysed during the current study are available in the [EMO-DB], [RAVDESS], [IEMOCAP] and [SAVEE] datasets.

- EMO-DB dataset: https://www.kaggle.com/datasets/piyushagni5/berlin-database-of-emotional-speech-emodb
- RAVDESS dataset: https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio
- IEMOCAP dataset: https://sail.usc.edu/iemocap/
- SAVEE dataset: http://kahlan.eps.surrey.ac.uk/savee/
- CREMA-D dataset: https://github.com/CheyneyComputerScience/CREMA-D

References

[1] A. A. Abdelhamid, E. S. M. El-Kenawy, B. Alotaibi, G. M. Amer, M. Y. Abdelkader, A. Ibrahim, and M. M. Eid, "Robust speech emotion recognition using CNN+ LSTM based on stochastic fractal search optimization algorithm," *IEEE Access*, vol. 10, pp. 49265-49284, 2022, https://doi.org/10.1109/ACCESS.2022.3172954.

[2] Z. Chen, J. Li, H. Liu, X. Wang, H. Wang, and Q. Zheng, "Learning multi-scale features for speech emotion recognition with connection attention mechanism," *Expert Systems with Applications*, vol. 214, p. 118943, 2023, https://doi.org/10.1016/j.eswa.2022.118943.

[3] A. Aggarwal, A. Srivastava, A. Agarwal, N. Chahal, D. Singh, A. A. Alnuaim, and H. N. Lee, "Two-way feature extraction for speech emotion recognition using deep learning," *Sensors*, vol. 22, no. 6, p. 2378, 2022, https://doi.org/10.3390/s22062378.

[4] S. Kumar, M. A. Haq, A. Jain, C. A. Jason, N. R. Moparthi, N. Mittal, and Z. S. Alzamil, "Multilayer Neural Network Based Speech Emotion Recognition for Smart Assistance," *Computers, Materials & Continua*, vol. 75, no. 1, 2023, https://doi.org/10.32604/cmc.2023.028631.

[5] S. Jothimani and K. Premalatha, "MFF-SAug: Multi feature fusion with spectrogram augmentation of speech emotion recognition using convolution neural network," *Chaos, Solitons Fractals*, vol. 162, p. 112512, 2022, https://doi.org/10.1016/j.chaos.2022.112512.

[6] D. Yang, S. Huang, Y. Liu, and L. Zhang, "Contextual and cross-modal interaction for multi-modal speech emotion recognition," *IEEE signal processing letters*, vo. 29, pp. 2093-2097, 2022, https://doi.org/10.1109/LSP.2022.3210836.

[7] J. Lei, X. Zhu, and Y. Wang, "BAT: Block and token self-attention for speech emotion recognition," *Neural Networks*, vol. 156, pp. 67-80, 2022, https://doi.org/10.1016/j.neunet.2022.09.022.

[8] B. Maji and M. Swain, "Advanced fusion-based speech emotion recognition system using a dual-attention mechanism with conv-caps and bi-gru features," *Electronics*, vol. 11, no. 9 p. 1328, 2022, https://doi.org/10.3390/electronics11091328.

[9] Y. Liu, H. Sun, W. Guan, Y. Xia, and Z. Zhao, Multi-modal speech emotion recognition using self-attention mechanism and multi-scale fusion framework," *Speech Communication*, vol. 139, pp. 1-9, 2022, https://doi.org/10.1016/j.specom.2022.02.006

[10] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, Self supervised adversarial domain adaptation for cross-corpus and cross-language speech emotion recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1912-1926, 2022, https://doi.org/10.1109/TAFFC.2022.3167013.

[11] S. Kakuba, A. Poulose, and D. S. Han, "Attention-based multi-learning approach for speech emotion recognition with dilated convolution,"

IEEE Access, vol. 10, pp. 122302-122313, 2022, https://doi.org/10.1109/ACCESS.2022.3223705.

[12] P. R. Prakash, D. Anuradha, J. Iqbal, M. G. Galety, R. Singh, and S. Neelakandan, "A novel convolutional neural network with gated recurrent unit for automated speech emotion recognition and classification," *Journal of Control and Decision*, vol. 10, no. 1 pp. 54-63, 2023, https://doi.org/10.1080/23307706.2022.2085198.

[13] R. K. Das, N. Islam, M. R. Ahmed, S. Islam, S. Shatabda, and A. M. Islam, "BanglaSER: A speech emotion recognition dataset for the Bangla language," *Data in Brief*, vol. 42, p. 108091, 2022, https://doi.org/10.1016/j.dib.2022.108091.

[14] B. B. Al-onazi, M. A. Nauman, R, Jahangir, M. M. Malik, E. H. Alkhammash, and A. M. Elshewey, "Transformer-based multilingual speech emotion recognition using data augmentation and feature fusion," *Applied Sciences*, vol. 12, no. 18 p. 9188, 2022, https://doi.org/10.3390/app12189188.

[15] J. Li, X. Zhang, L. Huang, F. Li, S. Duan, and Y. Sun, "Speech Emotion Recognition Using a Dual-Channel Complementary Spectrogram and the CNN-SSAE Neutral Network," *Applied Sciences*, vol. 12, no. 19, p. 9518, 2022, https://doi.org/10.3390/app12199518.

[16] J. Ancilin and A. Milton, "Improved speech emotion recognition with Mel frequency magnitude coefficient," *Applied Acoustics*, vol. 179, p. 108046, 2021, https://doi.org/10.1016/j.apacoust.2021.108046.

[17] H. A. Abdulmohsin, "A new proposed statistical feature extraction method in speech emotion recognition," *Computers & Electrical Engineering*, vol. 93, p. 107172, 2021, https://doi.org/10.1016/j.compeleceng.2021.107172.

[18] C. Zhang and L. Xue, "Autoencoder with emotion embedding for speech emotion recognition," *IEEE access*, vol. 9, pp. 51231-51241, 2021, https://doi.org/10.1109/ACCESS.2021.3069818.

[19] P. T. Krishnan, A. N. Joseph Raj, and V. Rajangam, "Emotion classification from speech signal based on empirical mode decomposition and non-linear features: Speech emotion recognition," *Complex & Intelligent Systems*, vol. 7, pp. 1919-1934 2021, https://doi.org/10.1007/s40747-021-00295-z.

[20] K. S. Chintalapudi, I. A. K. Patan, H. V. Sontineni, V. S. K. Muvvala, S. V. Gangashetty, and A. K. Dubey, "Speech Emotion Recognition Using Deep Learning," In *2023 International Conference on Computer Communication and Informatics (ICCCI)*, IEEE, Coimbatore, India, 2023, pp. 1-5, https://doi.org/10.1109/ICCCI56745.2023.10128612

[21] S Yildirim, Y. Kaya, and F. Kılıç, "A modified feature selection method based on metaheuristic algorithms for speech emotion recognition," *Applied Acoustics*, vol. 173, p. 107721, 2021, https://doi.org/10.1016/j.apacoust.2020.107721.

[22] G. Agarwal and H. Om, "Performance of deer hunting optimization based deep learning algorithm for speech emotion recognition," *Multimedia Tools and Applications*, vol. 80, pp. 9961-9992, 2021, https://doi.org/10.1007/s11042-020-10118-x.

[23] K. Manohar and E. Logashanmugam, "Hybrid deep learning with optimal feature selection for speech emotion recognition using improved meta-heuristic algorithm," *Knowledge-Based Systems*, vol. 246, p. 108659, 2022, https://doi.org/10.1016/j.knosys.2022.108659.

[24] Mustaqeem and S. Kwon, "Optimal feature selection based speech emotion recognition using two-stream deep convolutional neural network," *International Journal of Intelligent Systems*, vol. 36, no. 9 pp. 5116-5135, 2021, https://doi.org/10.1002/int.22505.

[25] S. Chattopadhyay, A. Dey, P. K. Singh, A. Ahmadian, and R. Sarkar, "A feature selection model for speech emotion recognition using clustering-based population generation with hybrid of equilibrium optimizer and atom search optimization algorithm," *Multimedia Tools and Applications*, vol. 82, no. 7, pp. 9693-9726, 2023, https://doi.org/10.1007/s11042-021-11839-3.

[26] S. Kanwal and S. Asghar, Speech emotion recognition using clustering based GA-optimized feature set," *IEEE Access*, vol. 9, pp. 125830-125842, 2021, https://doi.org/10.1109/ACCESS.2021.3111659.

[27] N. Barsainyan and D. K. Singh, "Optimized cross-corpus speech emotion recognition framework based on normalized 1D convolutional neural network with data augmentation and feature selection," International Journal of Speech Technology, vol. 26, no. 4, pp. 947-961, 2023, https://doi.org/10.1007/s10772-023-10063-8.

[28] C. Sun, H. Li, and L. Ma, "Speech emotion recognition based on improved masking EMD and convolutional recurrent neural network," *Frontiers in Psychology*, vol. 13, p. 1075624, 2023, https://doi.org/10.3389/fpsyg.2022.1075624.

[29] L. T. C. Ottoni, A. L. C. Ottoni, and J. D. J. F. Cerqueira, "A Deep Learning Approach for Speech Emotion Recognition Optimization Using Meta-Learning," *Electronics*, vol. 12, no. 23, p. 4859, 2023, https://doi.org/10.3390/electronics12234859.

[30] N. T. Pham, D. N. M. Dang, N. D. Nguyen, T. T. Nguyen, H. Nguyen, B. Manavalan, V. P. Lim, and S. D. Nguyen, "Hybrid data augmentation and deep attention-based dilated convolutional-recurrent neural networks for speech emotion recognition," *Expert Systems with Applications*, vol. 230, p. 120608, 2023,

[31] Z. Chen, J. Li, H. Liu, X. Wang, H. Wang, and Q. Zheng, "Learning multi-scale features for speech emotion recognition with connection attention mechanism," *Expert Systems with Applications*, vol. 214, p.118943, 2023.

[32] M. Khan, W. Gueaieb, A. El Saddik, and S. Kwon, "MSER: Multimodal speech emotion recognition using cross-attention with deep fusion," *Expert Systems with Applications*, vol. 245, p.122946, 2024.

[33] F. A. Dal Rì, F. C. Ciardi, and N. Conci, "Speech Emotion Recognition and Deep Learning: An Extensive Validation using Convolutional Neural Networks," *IEEE Access*, vol. 11, pp. 116638-116649, 2023.

[34] Y. Du, R. G. Crespo, and O. S. Martínez, "Human emotion recognition for enhanced performance evaluation in e-learning," *Progress in Artificial Intelligence*, vol. 12, no. 2, pp. 199-211, 2023.

[35] S. Debnath, P. Roy, S. Namasudra, and R. G. Crespo, "Audio-Visual Automatic Speech Recognition Towards Education for Disabilities," *Journal of Autism and Developmental Disorders*, vol. 53, no. 9, pp. 3581-3594, 2023.

[36] L. Pipiras, R. Maskeliūnas, and R. Damaševičius, "Lithuanian speech recognition using purely phonetic deep learning," *Computers*, vol. 8, no. 4, p. 76, 2019.

[37] A. Lauraitis, R. Maskeliūnas, R. Damaševičius, and T. Krilavičius, "Detection of speech impairments using cepstrum, auditory spectrogram and wavelet time scattering domain features," *IEEE Access*, vol. 8, pp. 96162-96172, 2020.

[38] Link for EMO-DB dataset: https://www.kaggle.com/datasets/piyushagni5/berlin-database-of-emotional-speech-emodb

[39] Link for RAVDESS dataset: https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio

[40] Link for IEMOCAP dataset: https://sail.usc.edu/iemocap/

[41] Link for SAVEE dataset: http://kahlan.eps.surrey.ac.uk/savee/

[42] Link for CREMA-D dataset: https://github.com/CheyneyComputerScience/CREMA-D

[43] M. M. Rahman and F. H. Siddiqui, "Multi-layered attentional peephole convolutional LSTM for abstractive text summarization," *Etri Journal*, vol. 43, no. 2, pp. 288-298, 2021.

Ramya Paramasivam

Ramya Paramasivam received her M. Tech. degree in Computer Science and Engineering and Ph.D. in Information and Communication Engineering from Anna University in the years 2005 and 2017 respectively. Her research areas are Artificial Intelligence and Wireless Communication. She is currently an Associate Professor in the Department of Computer Science Engineering, Mahendra Engineering College, for the last 6 Years. Dr.P.Ramya has over 19 years' experience in teaching, research and industry. She has taught many courses in Computer Science and Engineering at post graduate and undergraduate levels. Dr.P.Ramya has published and presented papers at many international journals/conferences / symposiums and delivered many invited / plenary / expert lectures in India.

K. Lavanya

K. Lavanya received the Ph.D. degree in Information and communication from Anna University, Chennai, India. She is currently an Assistant Professor in the department of Electronics and Communication Engineering, Velammal Engineering College, Chennai. She has a strong academic teaching and research experience of more than 18 years. Her research interests include Internet of Things, machine learning techniques, cryptography and network security.

Parameshachari Bidare Divakarachari

Parameshachari Bidare Divakarachari is currently working as a professor in the Department of Electronics and Communication Engineering at Nitte Meenakshi Institute of Technology, Bangalore, India; affiliated to Visvesvaraya Technological University (VTU), Belagavi, Karnataka, India. He has around 19 years of experience and has published over 200+ articles in SCI, SCOPUS, and other indexed journals and also in conferences. He serves as editorial board member, associate editor, academic editor, guest editor, and reviewer for various reputed indexed journals. He is also the founder chair for IEEE Information Theory Society, Bangalore chapter and IEEE Mysore Subsection. He is also the SAC Chair, IEEE Bangalore Section.

David Camacho

David Camacho (Senior Member, IEEE) received the Ph.D. degree (with Hons.) in computer science from Universidad Carlos III de Madrid, Getafe, Spain, in 2001. He is currently a Full Professor with Computer Systems Engineering Department, Universidad Politécnica de Madrid (UPM), Madrid, Spain, and the Head of the Applied Intelligence and Data Analysis Research Group, UPM. He has authored or coauthored more than 300 journals, books, and conference papers. His research interests include machine learning (clustering/deep learning), computational intelligence (evolutionary computation, swarm intelligence), social network analysis, fake news and disinformation analysis. He has participated/led more than 50 research projects (Spanish and European: H2020, DG Justice, ISFP, and Erasmus+), related to the design and application of artificial intelligence methods for data mining and optimization for problems emerging in industrial scenarios, aeronautics, aerospace engineering, cybercrime/cyber intelligence, social networks applications, or video games among others. He was the recipient of the best thesis award in Computer Science for his Ph.D. degree.