IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 42, No. 2, April, 2025, pp. 647-662
Journal homepage: http://iieta.org/journals/ts

A Combine CNN-RNN Based Approach for Augmenting the Performance of Speech Emotions Recognition

Nasir Sayed[1,2], Ghassan Husnain[1], Muhammad Shoaib[1], Yazeed Yasin Ghadi[3], Masoud Alajmi[4], Ayman Qahmash[5*]

[1] Department of Computer Science, CECOS University of IT and Emerging Science, Peshawar 25000, Pakistan
[2] Department of Computer Science, Islamia College Peshawar, Peshawar 25100, Pakistan
[3] Department of Computer Science and Software Engineering, Al Ain University, Al Ain 64141, United Arab Emirates
[4] Department of Computer Engineering, College of Computers and Information Technology, Taif University, Taif 21944, Saudi Arabia
[5] Department of Informatics and Computer Systems, King Khalid University, Abha 61421, Saudi Arabia

Corresponding Author Email: a.qahmash@kku.edu.sa

https://doi.org/10.18280/ts.420205

Received: 18 August 2024
Revised: 20 November 2024
Accepted: 10 April 2025
Available online: 30 April 2025

Keywords:
language understanding, features learning block, spectrogram, recurrent neural network, Speech Emotion Recognition (SER), long short-term memory

ABSTRACT

Due to the advancement of neural networks and the increasing demand for accurate and real-time Speech Emotion Recognition (SER) in human-computer interactions, it is necessary to compare existing methods and databases in SER in order to arrive at feasible solutions and a complete understanding of this open-ended problem in SER. To detect and recognize the emotions expressed in speech, various techniques have been used in the literature, including well-established speech analysis and classification techniques. These techniques, including speech analysis and classification, have been used to extract emotions from signals. In this study, we propose a novel method for analyzing signals called Wavelet-Scaled Spectrogram which combines the frequency and scale spectrum of a signal using wavelet transform. This method is effective in analyzing signals at different scales and frequency content. In order to train models for speech emotion identification, a large number of handcrafted features and intermediary depictions i.e., frequency-time plot that have traditionally been utilized in data compilation, collection, and analysis. The development of end-to-end models which extract characteristics and learn directly from raw speech signals to improve speech recognition has recently been studied by researchers following the emergence of deep learning. After training and evaluation on the famous speech databases EmoDB, RAVDESS and IEMOCAP, the proposed model is evaluated on various speakers in both speaker-independent and speaker-dependent modes and on a variety of different voices. When advanced preprocessing techniques or data augmentation are omitted from the proposed architecture, the results demonstrate that it can produce products comparable to those produced by the current state of the art. Three concurrent CNN pipelines and a series of modified local features learning blocks (LFLBs) achieved the highest classification accuracy outperforming some advance state-of-the-approaches.

1. INTRODUCTION

The crucial role inflammation plays in the beginning and development of coronary heart disease is well-known (CHD). However, the precise mechanism through which inflammation contributes to the pathophysiology of CHD remains unclear [1, 2].

Data usage becomes a more valuable asset as organizations move toward greater automation. Learning about new and more resource-efficient data generation and storage methods will benefit a wide variety of fields and organizations, enabling them to advance in their respective fields and organizations [1]. Recommendation engines and streaming services highly depend on the metadata associated with their material when making recommendations based on, among other things, the kind of content you have previously viewed on their platforms [2]. When data is gathered manually, annotation generation is time-consuming and costly, and the quality of the annotations varies significantly between contributors. Additionally, the availability of adequate metadata is a concern, as it has a detrimental effect on the quality of the generated recommendations. According to the research, superior metadata results in more accurate recommendations, which results in a more pleasurable user experience overall [3]. As the visual content of the video is the focus of the extraction process, a lot of emphasis is typically placed on it when metadata is collected from videos. On the other hand, streaming multimedia has rich audible data that can be used to extract the content's context, emotions, and other metadata [4]. Consequently, it is essential to recognize emotional expressions in speech while developing metadata and using all available content data to make smart decisions.

Traditionally, the identification and classification of emotions expressed in speech have been accomplished using intermediate representations such as various spectrograms, low-level descriptors [5], and parameters relating to various waveform properties such as mean frequency and fundamental frequency, and spectral slope. However, significant advancements in SER methods have occurred in recent years, with one particularly notable advancement being the incorporation of high-level descriptors and parameters relating to various waveform properties, such as fundamental frequency and spectral slope, into the algorithms. To successfully implement these features, it is frequently necessary to have expert domain knowledge and resources, just as entering metadata to engineer these features manually. Deep Learning has outperformed traditional methods in a variety of applications due to the development of neural network architectures such as Convolutional Neural Net (CNN) and innumerable types of Recurrent Neural Networks (RNN) [6]. Networks are now able to learn from the raw audio stream itself thanks to deep learning [7]. Since manual feature creation is no longer necessary, time and money are conserved. Deep learning based CNNs can train and extract features directly from unprocessed acoustic signals, eradicating the requirement for overpriced and time-consuming human feature engineering processes [8]. In this study explores the possibility of enhancing the ability to recognize speech emotions by combining equivalent CNNs for chin extraction with Short Long-Term Memory (LSTM) nets for classification SER. With the help for this inquiry, the unprocessed audio signal waveforms will be studied. The primary objective of this study is to investigate a hybrid deep neural network that predicts the emotional content of speech using machine learning methods and learning from raw audio. Due of their dynamic nature and the fact that many people have trouble accurately articulating emotions, emotions are challenging to recognize. de Lope [9] claims that human listeners, for instance, can detect the sentiment of an unfamiliar speakers with 60% an accuracy, which is roughly five periods the chance baseline accuracy for the tests under discussion. In addition, there is a lack of training data which Khan et al. [10] remark to as well as the fact that classic feature-based models typically encounter implementation issues making it difficult to acquire emotional qualities from low-level speech signals. Although deep neural networks with convolutional layers are capable of handling high-dimensional input [11, 12], solving manual feature engineering challenges introduces bias into the feature selection process necessitating the use of domain-specific expertise from professionals to design features [8]. High-dimensional data can be processed by convolutional deep neural networks, which can also automatically train features that are impervious to small imperfections and variations. Several studies have demonstrated that deep neural network approaches extract more accurate feature representations and outperform conventional techniques for example Hidden Markov Models (HMM) and Gaussian Mixture Model (GMM) [13], compared to other methods, DNN shave a lower sensitivity to minor changes in the input features and a higher level of robustness in SER when exposed to changes in the speaker's voice, environment, and bandwidth [14]. As a result, DNNs generalize better than other network types, such as external networks. Peer convolutional layers are incorporated into our deep learning architecture for multitemporal, feature removal and temporal of long-term modelling, which allows us to reduce feature engineering difficulties while retaining an effective and simple pipeline [15-17].

It is particularly advantageous for evaluating deep learning architectures because of being viewed as a static or dynamic classification problem, allowing for applying a wide variety of different modeling approaches [18]. In comparison to static modeling, which seeks to recognize emotions across an entire utterance, dynamic modeling is frame-based and seeks to recognize emotions within each frame [5].

The results of this study are anticipated to be helpful to market participants interested in creating end-to-end SER models or, among other things, improving the performance of current models. The creation of models for video classification and metadata production, to mention a couple, are some examples of potential applications. With this knowledge, it might be able to choose the best method for include the audio component to optimize the information in the video. This might be used as a stand-alone model or additional probably this is a part of a group model with extra modalities depending on the outcomes of this research. The main contribution of the proposed model can be summarized as:

Develop a speech-based user emotion recognition system as the primary outcome of the proposed research.

Introduce an interactive approach utilizing the Wavelet-Scaled Spectrogram method for the analysis of original speech. This method integrates the frequency and scale spectrum of the speech signal through wavelet transform, facilitating the extraction of valuable insights and the reconstruction of 2D RGB image data for subsequent analysis and interpretation.

Construct a custom features learning block based on Convolutional Neural Network (CNN) layers. The CNN features extracted are then fed into a custom Long Short-Term Memory (LSTM) layer-based classification model designed for the task of SER.

Evaluate the robustness and generalizability of the proposed model by assessing its performance on a variety of voices and under different conditions. This objective aims to investigate the model's adaptability to diverse speech patterns and environmental factors, ensuring its efficacy in real-world applications.

Train and validate the proposed model using two benchmark speech signal datasets, namely EmoDB, RAVDESS, and IEMOCAP. This evaluation is conducted across various speakers in both speaker-dependent and speaker-independent modes, providing comprehensive insights into the model's performance.

The study introduces a novel Wavelet-Scaled Spectrogram method, which effectively integrates the frequency and scale spectra of speech signals using wavelet transform, enabling precise feature extraction and reconstruction of 2D RGB image data for analysis. This surpasses traditional methods such as Mel-Spectrograms and Short-Time Fourier Transforms by capturing richer temporal and spectral details, particularly in complex emotional contexts. The authors should clarify how this approach uniquely enhances SER by addressing limitations of scale and frequency content analysis in existing techniques.

The proposed CNN-LSTM hybrid architecture advances the state-of-the-art by combining CNN's ability to extract high-dimensional spatial features from raw audio with LSTM's strength in modeling temporal dependencies. Unlike conventional methods requiring manual feature engineering and domain expertise, this automated feature extraction pipeline minimizes preprocessing, reduces implementation

biases, and significantly improves robustness to variations in speaker voices and environmental noise. The key advantage is its generalizability across diverse datasets and emotional expressions, demonstrated through evaluations on benchmark datasets like EmoDB, RAVDESS, and IEMOCAP.

The study further distinguishes itself by proposing a custom feature extraction block that reduces raw audio dimensionality while retaining essential information for classification. This block, paired with LSTM layers and optimized hyperparameters (e.g., Adam optimizer), ensures efficient and accurate emotion recognition. The authors should emphasize how these design choices address challenges such as limited training data, sensitivity to input variations, and complexity in feature engineering.

Below is the major contrition of the proposed work:

- The study introduces a novel method for analyzing speech signals, the Wavelet-Scaled Spectrogram, which effectively combines the frequency and scale spectrum using a wavelet transform. This approach provides a powerful tool for extracting insights and information from speech signals, surpassing traditional methods in terms of scale and frequency content analysis.
- The proposed model integrates three concurrent CNN pipelines with Long Short-Term Memory (LSTM) units, leveraging the strengths of both CNNs and RNNs in feature extraction and temporal sequence modeling. This architecture enhances the accuracy and robustness of SER, particularly in complex emotional contexts.
- The research introduces a unique feature extraction block that reduces the dimensionality of raw audio signals while retaining essential information for classification. The classification block, inspired by state-of-the-art methods, is simplified yet powerful, comprising LSTM, Fully-Connected (FC), and Softmax layers for efficient emotion classification.
- The model's effectiveness is rigorously evaluated on three prominent SER datasets-EmoDB, RAVDESS, and IEMOCAP-ensuring the generalizability and robustness of the proposed approach across diverse emotional expressions and languages.
- The study thoroughly explores and optimizes key hyperparameters, including learning rates, pooling strategies, and optimization algorithms. The use of the Adam optimizer, in particular, led to significant improvements in convergence speed and overall model accuracy.
- By minimizing pre-processing requirements and excluding data augmentation, the study demonstrates that the proposed model can achieve high performance with limited manual intervention, making it more accessible and applicable in real-time scenarios.

The article begins with an **Introduction** that outlines the significance of SER in human-computer interaction, the challenges in the field, and the key contributions of this study, including the development of the Wavelet-Scaled Spectrogram and a combined CNN-RNN model. The **Literature Review** covers existing SER techniques, the use of wavelet transforms in signal processing, and the role of CNNs and RNNs in emotion recognition, alongside an overview of key datasets like EmoDB, RAVDESS, and IEMOCAP. The **Methodology** section details the proposed model, including the feature extraction and classification blocks, the datasets used, pre-processing steps, and hyperparameter tuning. Finally, the **Results** section presents the performance evaluation of the model across various datasets, compares it with state-of-the-art methods, and discusses the generalizability and implications of the findings.

2. LITERATURE REVIEW

To be more precise, developing a model for emotion recognition will be the primary objective of this investigation. As a result, we will begin our investigation by examining the realm of emotions. Sentiments can be discrete or continuous, depending on how the individual experiencing them perceives them. Emotions can be classified according to their linguistic origins, and categorical interpretation is the study of expressions that convey a variety of mental states or emotions in a variety of situations [19]. The term universal emotions coined by evolutionary psychologist Charles Darwin, refers to feelings elicited consistently throughout the world, regardless of location. Darwin proposed it for the first time in 1859. The subsequent popularization of this concept resulted in the emergence of Ekman's discrete collection of fundamental emotional states [20], which includes the emotions of happiness and sadness and fear, surprise, anger, and disgust. According to Ekman's findings, the quantity of emotions listed above, dubbed The Big Six were separately distinguishable and encompassed the entire emotional space. This group of fundamental emotions is frequently mentioned in the literature, and some variations, extensions, and subsets of these emotions can be observed in action.

2.1 Speaker dependent vs speaker independent

Typically, speech recognition systems (SRS) are classified into two categories: those that are speaker-independent and those that are speaker-dependent. The term "dependent speech recognition systems" refers to speaker-dependent ones. Independent speech recognition systems operate independently of the speaker's voice. Prior to implementing any speaker-dependent system, each intended user must be trained to operate it. Alternatively, speaker-independent systems are introduced on many utterances of all vocabulary items, allowing for increased flexibility. The results are the most effective when hundreds of people collaborate on a project [21, 22]. To accomplish this, substantial resources must be allocated, and the effective collection of data. The research of Liang [23], who found an average recognition 60% of accuracy for human listeners recognizing an unfamiliar speaker's emotion, shows that it is a challenging task for humans as well. The reported human accuracy is around five times that of random baseline accuracy, according to a survey of about 30 research from the 1980s [24]. EmoDB and RAVDESS two cutting-edge datasets that are publicly accessible online, are described and given a brief overview. Then, to help us create our suggested architecture, we perform a literature review. Numerous LSTM subtypes, the relationship among kernel and pooling size, the notion of native feature learning blocks and parallel multi-temporal CNNs layers are among topics we address.

2.2 EmoDB Dataset

Abdusalomov et al. [25] pioneered the concept of capturing an audio signal's multitemporal characteristics using a series of Spectro temporal box filters, which were not introduced until 2010. (STBFs). In information theory, a single time-scale combination of information spans three distinct time scales is

referred to as the STBF. Low-level features like MFCCs and LPCCs can be recorded and evaluated because of the small-time scale utilized [26]. The local derivatives of those features are computed on a medium-timescale basis in accordance with the time scale of the low-level features. The results from the medium timescale are also condensed into a single summary statistic known as the long timescale, which is defined as a duration longer than one year. Combining these three time periods, the researchers [19, 25] discovered a speaker-independent precision of 77.12% on EmoDB, which they shared with the general audience.

To develop affect-salient features for SER, Rezapour Mashhadi and Osei-Bonsu [27] suggested a semi-CNN in 2014. To learn these features, the semi-CNN was trained on spectrograms with two different resolutions as the input. The semi-CNN is used to generate affect-salient features after training in two stages, first unsupervised and then semi-supervised, and is then input into a linear SVM for SER, which is then trained in a second stage to generate SER features, and so on. After deep evaluation, the model for EmoDB demonstrated that a semi-CNN configuration was resistant to distortion brought on by the environment and the speaker when assessed without utilizing a speaker, achieving an accuracy of 85.20.445 percent. Another study [28], released at the end of 2015, looked at the application of unique wavelet packet (WP) features for SER that included tree pruning and filter banks to boost the efficiency of conventional MFCC-based methods. To improve SER performance, the authors of this study looked at the usage of novel WP features that integrated filter banks and tree pruning. The results showed that the suggested model, employing six different emotional classes from the EmoDB, had an average precision of 75.51 percent.

Figure 1 visually represents the proposed CNN architecture for SER, utilizing Spectrogram RGB images to extract rich spatial and temporal features, enabling the model to effectively capture emotional nuances in speech signals. The researchers presented a simple CNN architecture for extracting salient discriminative features from spectrograms with excellent accuracy. They claimed that this design could accurately extract substantial discriminative features from spectrograms [29]. The architecture depicted in Figure 2 has three 2D Conv layers, three dense layers, one probability layer, and three fully connected layers. Two dropout layers with a 50 percent dropout ratio were added in the design to minimize overfitting. Each audio sample in the training data was split into many chunks, and then FFT was used to determine the frequencies at various locations in the sound for each chunk. The model was supplied the resulting number of spectrograms for each audio sample, then fed back into it as described previously. We noticed throughout our research that the pre-trained AlexNet model performed poorly in this scenario and that transfer learning had no effect on learning performance in this context. When trained and tested on the Berlin EmoDB [30], the newly trained CNN performed brilliantly, attaining an overall test accuracy of 52 percent across all seven categories across all training and testing sessions. The newly trained CNN also displayed excellent adaptability, earning a total test accuracy of 52 percent across all seven categories across all training and testing sessions. However, researchers claim that the model failed poorly when discriminating between fear and enjoyment. 0.71 points were lost in total, 0.71 in the training set, and 0.95 in the test set, for 0.71 points lost in the game. According to the researchers, the dataset generated over 3000 spectrograms, generating around 500 images per emotional state for each emotion [31]. While discussing their desire to deploy the models for emergency phone calls, the authors bring out numerous critical problems, including background noise, poor transmission quality, and the Lombard effect. Numerous these factors also apply to video applications since, in many circumstances, when background noise is present in the recording, a clean voice signal cannot be recovered, underscoring the importance of several of these considerations. Satt and colleagues showed that when only convolutional deep neural networks with minimal complexity were used, they reached a 66 percent accuracy over four emotions [32]. On the same evaluation set, a mixed convolutional model with a greater level of complexity attained a prediction accuracy of 68 percent, the same as in 2017. Although the spectrogram's frequency resolution of 40Hz yielded the highest level of accuracy, there was a decrease of approximately 4 to 7 percentage points in accuracy when the resolution was increased to 60Hz. The researchers employed a series of overlapping Hamming windows with window sizes of 60ms and shifts of 20ms, where the window sizes overlapped by 60ms and the window shifts were equal to a 20ms window shift. This study also discusses how raw spectrograms can be used to define speech characteristics and how they allow for effective non-speech background management, even at noise levels that are comparable to the strength of the speech signal. To deal with background noise properly and efficiently, harmonic filtering is required to manage 'signal 'to noise ratios as great as 1:1 [33]. As a follow-up to our earlier discussion, we will evaluate some prior research conducted utilizing the RAVDESS dataset.

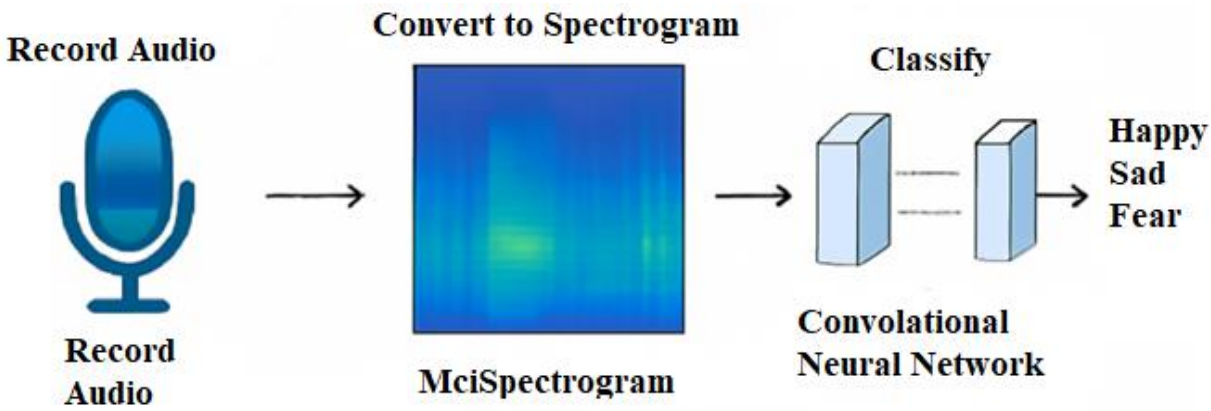

Figure 1. Spectrogram RGB images based proposed CNN architecture for Speech Emotions Recognition [15]

2.3 RAVDESS

The challenge in SER, as identified by Dutt and Gader [34], lies in the extraction of speaker-independent features that effectively capture emotional information conveyed through speech and transmit it to the user. To extract manually constructed features from the data, techniques such as continuous wavelet transforms (CWT) and prosodic coefficients including RMS energy, ZCR, and entropy are necessary. A support vector machine (SVM) classifier was trained using these features to address the issue. The Quadratic SVM model achieves an overall accuracy of 60.1 percent when subjected to a 5-fold cross-validation process on the RAVDESS dataset.

Singh et al. [35] proposed a deep learning model capable of handling multiple acoustic classification tasks by leveraging shared features across speaker identification, accent recognition, and emotion detection. The model outperformed several task-specific models that had previously been tested in the lab, with accuracy of 64.4 percent on RAVDESS for emotion recognition (ER). Deep Residual Networks and a gate mechanism were used to generate the model (ResNets).

MFCCs Chromagrams, Mel-scale spectrograms, Tonnetz representations, and audio spectral contrast features are only a few examples of the five different spectrum-based features that Isa and colleagues [36] presented as the foundation of a CNN-based architecture. The accuracy rates of the researchers' model were found to be 71.61 percent, 86.1 percent, and 64.3 percent respectively when tested using RAVDESS, EmoDB, and IEMOCAP. After 700 iterations, the model for RAVDESS achieves its maximum accuracy of 71.61 percent, which is the best accuracy feasible under the given conditions. Whether the authors [37] use of fivefold cross-validation on the whole datasets or random data partitioning renders the classification speaker-independent is arguable, however it is inconsistent with explanations in other state-of-the-art literature [38]. Then, to better comprehend the possibilities of LSTMs, we will look at a number of distinct types and how they are used in networks.

2.4 lstm based RNN

The end-to-end method proposed by Tellai et al. [39] was successful, as shown by their analysis of a convolutional RNN coupled with two bi-directional LSTMs (128 total units). The model is trained using 9600000-dimensional input vectors made up of 6 second raw waveform signals since the RECOLA dataset [40]. Based the input vector the model predicts three-dimensional emotions in the alerting and sentiment regions. The accuracy of the model is assessed by computing the error rate of the objective function, which is then used to assess the model's general performance. The RECOLA platform outperformed previously conventionally designed features, according to the findings of researchers Trigeorgis et al. [41] found that the network's overall performance was comparable to the unidirectional approach even though the final model contained two bidirectional LSTM layers. When bidirectional LSTMs were fed frames of SAR spectrograms, Fayek et al. [42] found a similar result: future context contributes very little to the network's performance.

In summary, the literature review highlights advancements in SER through the shift from traditional feature-based methods like MFCCs to deep learning approaches using CNNs and LSTMs. Studies on benchmark datasets, including EmoDB and RAVDESS, showcase improvements in accuracy and robustness by leveraging wavelet transforms and spectrogram-based architectures. Speaker-dependent and independent systems are discussed, emphasizing challenges like background noise and limited data. Recent hybrid CNN-LSTM models address these issues by combining high-dimensional feature extraction with temporal modeling, achieving superior performance and generalization over traditional and task-specific methods.

3. METHODOLOGY

The proposed model, which comprises of two main building blocks, the primary one is feature extraction and another one is classification block which is described in this section of the paper. A brief description of the proposed network architecture is also presented. The two primary building blocks of the network architecture proposed in this paper are an extraction block for features and a classification block for classification. Later in this chapter, the datasets and resources which are used to develop the network are discussed in greater detail. Additionally, considerable attention is paid to how the information is pre-processed. After the paper, we discuss the chosen hyperparameters, the rationale for their selection, their values as well as we provide a brief overall view of the evaluation procedure. Figure 2 presents a comparative analysis of machine learning and deep learning approaches for SER, highlighting the superior feature extraction and classification capabilities of deep learning methods over traditional machine learning techniques.

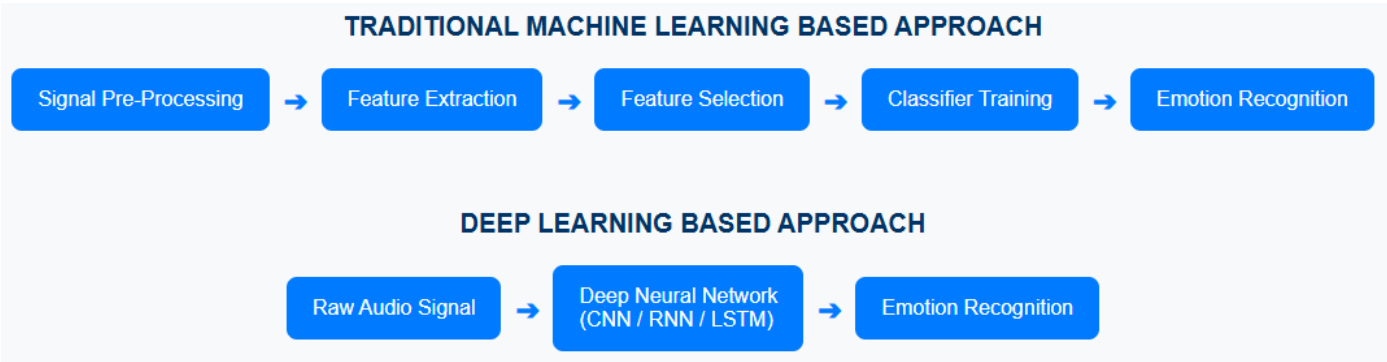

Figure 2. A comparison of machine learning and deep learning approach for SER [6]

3.1 Wavelet-Scaled spectrogram

The proposed Wavelet-Scaled Spectrogram method combines the frequency spectrum of a signal as it varies with time and the scale spectrum of a signal, which gives information about the frequency content of the signal at different scales. The method is obtained by applying a wavelet transform to the signal, computing the power of the wavelet coefficients, and plotting the power as a function of time and scale. This representation allows for the analysis of signals at different scales, while also providing information about the frequency content of the signal. The wavelet function used in the transform can be any wavelet function such as Morlet wavelet, Haar wavelet, etc. It is important to note that the proposed method provides a valuable tool for extracting insights and information from signals that cannot be obtained by traditional methods.

The Wavelet-Scaled Spectrogram method provides a significant advantage over traditional representations like spectrograms and Mel-spectrograms by leveraging the multi-scale nature of wavelet transforms. Unlike fixed-resolution methods, wavelet transforms offer variable window sizes, enabling precise temporal resolution for high-frequency components and better frequency resolution for low-frequency components. This adaptability captures subtle and dynamic emotional patterns in speech that are often missed by traditional methods. Additionally, wavelet transforms effectively handle the non-stationary nature of speech signals, preserving critical emotional cues. The proposed approach integrates these advantages, creating rich, context-aware representations that enhance accuracy and generalization in SER.

Proposed Image Generation Algorithm

Input: time series signal $x(t)$

Initialize the wavelet function, $W(t)$

Define the sceale and translation parameters a,b

Define the wavelet transform function:

$$c(a,b) = \left(\frac{1}{a}\right) * \int x(t) * W^{*}\left(\frac{t-b}{a}\right)dt$$

Compute the power of the wavelet coefficients, $|c(a,b)|^2$.

Plot the logarithm of the power of the wavelet coefficients as a function of time and scale to obtain the Wavelet-Scaled spectrogram

Output: Wavelet-Scaled spectrogram

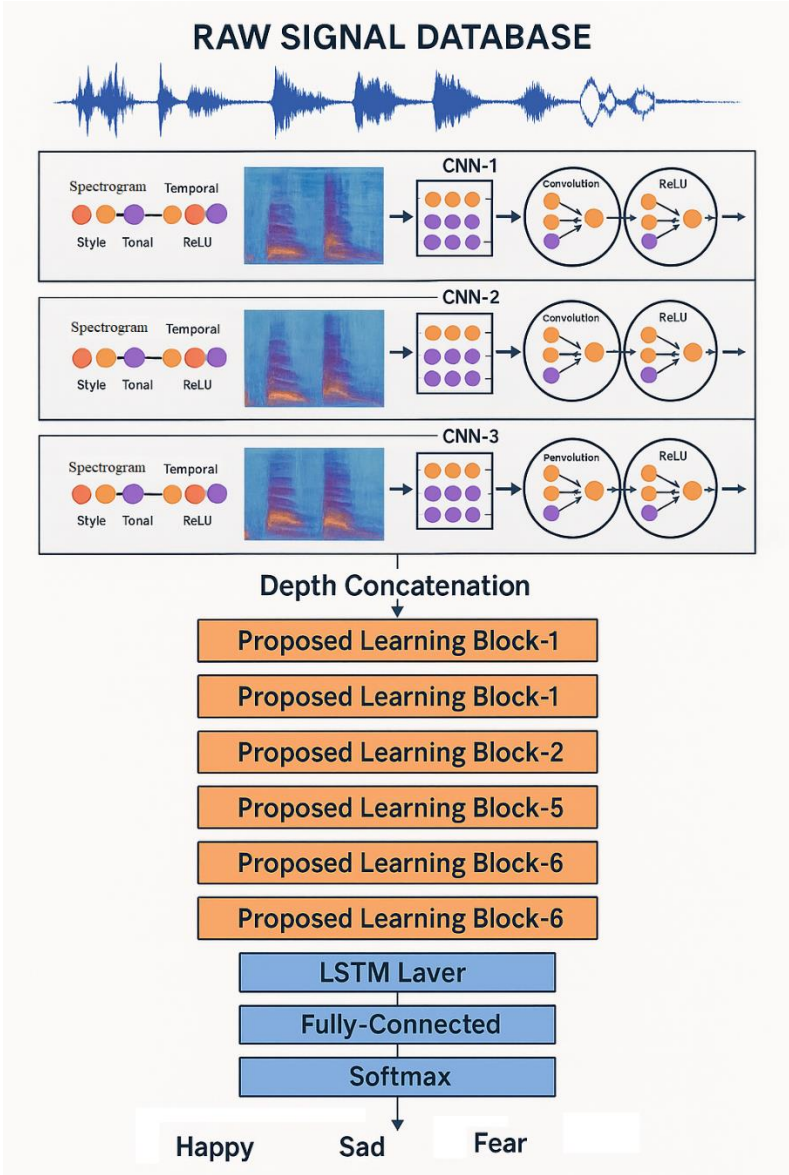

Figure 3. Proposed LSTM-RNN architecture using the LFLBS Temporal information for emotion recognitions

3.2 Network architecture

A hidden-to-out put function, and an input-to hidden function is a few of the methods for building LSTM-based deep neural networks [43]. Gunasekaran et al. [44] demonstrated that the deep input-to-hidden strategy for achieving high performance could be implemented in speech recognition. DNNs are represented at a higher level of abstraction in this strategy to extract speech features that are then fed into RNNs for speech recognition. As a result, the suggested architecture integrates this kind of input and the concealed LSTM technique (as depicted in Figure 3, which provides a summary of the proposed network architecture). Additionally, the parallel CNNs multitemporal structure of Nfissi et al. [45] and the LFLB feature extraction units of are incorporated into the network's suggested design. The proposed network's two critical building blocks are an extraction block for features and a classification block for classification. These are the two critical building blocks of the network that we propose. Parallel convolutional layers (PCL) separate the speech into three different temporal resolutions in the feature extraction block; these three resolutions are then combined and fed through a sequence of LFLB units these LFLB units extract the essential features and to reduce the resolution of the classification block representation; and finally, the classification block is composed of parallel convolutional layers that extract t SoftMax is a simple classification block that, once the network is fully connected, generates network's outputs categorical representations after the process. Three layers constitute the classification block LSTM layer, FC layer and SoftMax layer. The most advanced in them is the LSTM layer. For the learning process to be as effective as possible, both blocks must operate together.

Figure 3 depicts the proposed LSTM-RNN architecture, which leverages the Local Feature Learning Block with Temporal Information (LFLBS) to enhance the recognition of emotions by effectively capturing sequential dependencies in speech signals.

3.2.1 Feature extraction block

This block's main role is to take the functional characteristics of a raw signal and store them in the processor's memory. Numerous advantageous features of the model contribute to the accurate classification of previously unknown data, thereby increasing the model's overall predictive power. To represent the input raw audio signal at a sampling rate of 16 kHz as a vector of 128000 bits, the input vector must be the same length as the input raw audio signal. The audio vector's dimension must be reduced to ensure that the classification block and the LSTM learn as efficiently as possible. Due to the feature extraction process, the signal's dimension can be reduced, which can be accomplished via strides or pooling. When using strides, the maximum amount of information is considered; when using max pooling, the maximum amount of information is extracted, and the less significant pieces of information are ignored. According to Tang et al. [46], the overlap rate (R) should be kept as low as possible to achieve an R of 0.5.

Additionally, they discovered that the maximum pooling technique outperforms overstrides when decreased dimensionality. When comparing different pooling strategies, several researchers, including Liang et al. [23] and Nfissi et al. [45], discovered that maximum pooling outperforms all other pooling strategies. We will compare two layers max-pooling and average-pooling and determine whether layer is superior for the basic design of the suggested model based on the uncertainty in the comparisons' specifics in both studies. We split the feature extraction block into two halves, each of which has parallel convolutional layers and LFLBs sequences.

3.2.2 Feature extraction block

In comparison to other classification blocks, this one only includes three layers: LSTM (FCL), and a Softmax layer. The classification block was inspired by Zhao et al. [47]. We developed and implemented a single unidirectional LSTM unit in response to the findings of several previous studies, which will have a negligible effect on the network's overall performance in the future. Our experiments included testing the LSTM with 64, 128, and 256 cells, modulating the quantity of cells within the LSTM architecture while simultaneously tracking the level of accuracy. The architecture of the proposed features learning block is illustrated in Figure 4, highlighting its capability to learn discriminative features essential for emotion recognition.

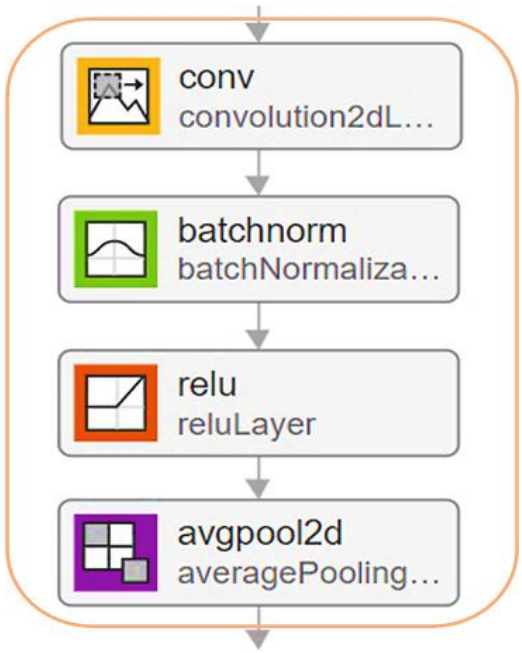

Figure 4. Proposed features learning block

Their initial experiments discovered that using 1024 units for the FC layer resulted in an excellent performance. These findings corroborated those of Nfissi et al. [45], whose findings served as the motivation for the investigators' parallel multitemporal CNN architecture. According to Nfissi et al. [45], it is intended for use in discriminative representation learning, and the FC layer is purpose-built for this purpose.

3.3 Dataset

In order to validate the effectiveness of the proposed model in speech-based emotion recognition, rigorous evaluation was conducted using three prominent datasets: The EmoDB [48], The RAVDESS [49], and the IEMOCAP [28]. These datasets were chosen for their comprehensive coverage of diverse emotional expressions in speech. The utilization of multiple datasets ensures a robust evaluation and provides insights into the generalizability and performance of the proposed model across different emotional contexts.

3.3.1 The EmoDB

Berlin EmoDB was chosen as the starting point for this investigation because it is a well-known corpus within the SER created in 2005 and has been the subject of previous investigations. There are 535 utterances in German divided into seven categories, each corresponding to one of the dataset's emotions: anger, contentment, unhappiness, anxiety, unbiased, disgust, and tediousness. The dataset contains 535 German utterances classified into seven categories, each corresponding to one of the emotions. The dataset can identify a extensive variety of emotions. Among the emotions that can be experienced are anger, unhappiness, anxiety, neutrality, contentment, disgust, and tediousness. In five of Ekman's Big Six emotions are present, as is the emotion of boredom as discussed in section 2.1.1. Figure 5 demonstrates the schematic overview of how this occurred. Unlike the RAVDESS dataset, the EmoDB utterances are simulated; however, their content is more diverse and drawn from everyday communication than the RAVDESS dataset. The EmoDB Dataset contains a relatively even distribution of samples across the various classes, except the category "disgust," which contains only 46 samples. The EmoDB Dataset contains a relatively even distribution of samples across the various classes. In the EmoDB Dataset total number of samples is reasonably evenly distributed across the various classes of models.

3.3.2 The RAVDESS

In this investigation, the second dataset analyzed was RAVDESS, short for Ryerson Audio-Visual Database of Demonstrative Song and Speech. Ryerson University created RAVDESS, a collection of audio and video song recordings and emotional speeches. Both of the given below expressions have been recorded 1440 in this collection.

"Boys are playing football in the street."

"A man is ringing the doors bell."

Twenty-four actors are delivering these statements (12 females and 12 males), with each statement accompanied by an expression representing one of eight distinct emotions. This dataset includes the emotion "calm," which is absent from the Ekman dataset, and it allows for the detection of five of “Ekman's” Big-Six Emotions, as well as the emotion sadness, which is also absent from the Ekman dataset. Ten annotations were made to the collection of recordings, using the terms emotional validity, intensity, and genuineness to differentiate them. Three categories of annotations were established: According to the results, human accuracy was measured to be 62% on the RAVAILONEDESS dataset for all intensities (average and strong), as well as for audio-only detection. Currently, the simulated database contains only North American accents and it is constructed by recording and analyzing professionally trained actors' vocal portrayals and emotional expressions.

Table 1. Ekmans Big-Six emotions schematic overview and emotions existing in EmoDB and RAVDESS

Ekman Big Six	EmoDB	RAVDESS	IEMOCAP
Calm	Happiness	Happiness	Happy
Disgust	Disgust	Disgust	Sad
Sadness	Fear	Fear	Neutral
Surprise	Happiness	Boredom	Angry
anger	Disgust	Disgust	
Fear	Anger	Anger	Excited
Happiness	Sadness	Sadness	

Table 1 provides a schematic overview of Ekman’s Big Six emotions, illustrating their mappings to the existing emotion labels in the EmoDB, RAVDESS, and IEMOCAP datasets. The table highlights variations in emotion labeling across these datasets, emphasizing the need for normalization in multi-dataset analyses.

3.3.3 The IEMOCAP

The IEMOCAP (Interactive Emotional Dyadic Motion Capture) dataset is a multimodal dataset that contains videos, audio recordings, and transcriptions of dyadic interactions between two actors. The dataset was created by researchers at the University of Southern California (USC) and contains a total of 10 hours of recordings of actors in five different emotional states: neutral, happy, sad, angry, and excited. The dataset includes a total of 12 actors (6 female, 6 male) and contains a total of 5,080 utterances in total. The data was recorded in a controlled environment, and the actors were asked to perform in a variety of scenarios such as telling a story, giving a presentation, or having a conversation. The IEMOCAP dataset is widely used in research related to SER, natural language processing, and human-computer interaction.

3.4 Pre-processing

To ascertain how much feature extraction can be further done to the model, adhering to the fewest possible pre-processing requirements is critical. We generated the data to train the models using load function Librosa's, with 16kHz sampling frequency and a sampling rate of 512 samples per second. Different audio libraries for the Python programming language were investigated, including Essentia and SciPy, and it was discovered that the audio library used did not affect the amount of time required to train the neural network. As described previously, a one-dimensional vector of floating-point values is created by sampling an audio file and storing the results in a one-dimensional vector. Because each audio file has a unique volume level, we normalize the signal values using the root-mean-square energy (RMSE), which is follows:

$$Error = Actual - Predicted \quad (1)$$

$$SE = (Error)^2 \quad (2)$$

$$MSE = 1/n \sum_{i=1}^{n} SE \quad (3)$$

$$\text{RMSE} = \sqrt{MSE} \quad (4)$$

X2i denotes i-th sample energy of the audio signals whereas “n” denotes the signal's length (n is the number of bits). Because the audio vectors are generated from audio files, their lengths will vary depending on the audio file used as an input source. This requires us to pad or offset the input vector to ensure that all inputs are of the same length, which is 8 seconds of audio. A random amount is added to the ends of audio vectors with lengths less than 128000, and a random amount is subtracted from the ends of audio vectors with lengths greater than 128000. The practice of data augmentation will be excluded from the scope of the investigation based on the findings of this study. Possessing the ability to train on a more extensive set of data points, even if those data points are generated in a virtual environment, is typically associated with developing more accurate models. As a result, it was determined that data augmentation would not be necessary for

this instance because the study would require only minimal manual pre-processing. Due to the hardware available capacity limitation, it was determined to exclude augmentation of data from further consideration for both the IEMOCAP and MSP-IM Providers datasets.

3.5 Parameter tuning

Liang et al. [23] discovered that maximum pooling performed the best, and they published their findings in this journal. This paper presents an abbreviated version of the proposed architecture, which is intended to evaluate various pooling strategies to determine which is the preferred option when choosing between the two primary pooling strategies: maximum pooling and average pooling. This enables us to determine which pooling strategy is the most advantageous for our particular circumstances. The initial section of the feature extraction block contains a single convolutional pipe used to evaluate pooling in the straightforward architecture for pooling evaluation. Two convolutional pipes make up the remainder of the feature extraction block. The process is completed by the convolutional pipes that comprise the remainder of the feature extraction block. Given the presence of BN in every LFLB in the network, it is unsurprising that the initial learning rate is non-significant [18]. The optimal initial learning rate was determined after preliminary experimentation, with learning rates ranging from 0.1 to 0.00001. It was possible to determine whether or not an optimization had occurred by examining the precision and time to convergence. Early stopping was used throughout the training process to minimize the risk of overfitting. As a result, after a predetermined number of epochs or patience, we stop processing the data. After evaluating various optimization algorithms, including Stochastic Gradient Descent (SGD) [50], AdaMax [51], and RMSprop, a more comprehensive study was conducted. The optimizer chose the final option, Adam, because of its high accuracy, rapid convergence and smooth learning curve during the initial phase of experiment, all of which the optimizer desired.

3.5.1 Final hyper parameter

In Table 2 below grants the hyper parameters last outline based on given literature and experimental results. The hyperparameter tuning process involved exploring a range of values for key parameters, guided by insights from existing literature and iterative experimentation. For learning rate, values were tested between 0.001 and 0.1 to balance convergence speed and model stability, with 0.1 yielding the best results. Batch sizes ranging from 8 to 64 were evaluated, and a size of 16 was selected for optimal performance and computational efficiency. The number of epochs was finalized at 50 based on performance saturation observed during training. The Adam optimizer was chosen due to its adaptive learning capabilities, which enhanced convergence. The final configuration, summarized in Table 2, reflects the optimal parameters determined through this systematic tuning process.

Table 2. Final Hyper parameters

Type	Parameter
Epochs	50
Base Learning Rate	0.1
Optimizer	Adam
Batch Size	16

The best configuration was selected through iterative testing of hyper parameters, including learning rate (0.001–0.1), batch size (8–64), and epochs, with 0.1, 16, and 50 yielding optimal results. The Adam optimizer was chosen for its adaptive learning capabilities and robust convergence.

3.5.2 Evaluation

To verify the results, a three-fold cross-validation procedure is used during the evaluation process. To begin, a random training riven of 20% is used for the initial speaker-dependent analysis is conduction. This clearly means that the “model is trained” and calculated on various data points but they can all come from the similar speakers. The model is then evaluated using data points collected from the same speaker who provided the input data. As a result, this initial assessment can be viewed as highly subjective and highly dependent on the individual providing it. Following that, an independent evaluation of the proposed model is conducted. To accomplish this, we remove two speakers from the training set prior to training and evaluate our model using data points generated by these two speakers. Our validation set includes one female and one male speaker who were not included in the training set, which allows us to be less reliant on the gender of our training set speakers. This is done to avoid becoming dependent on someone's sexual orientation. To ensure that our validation set contains a sufficient number of data points with a sufficient amount of variation to conduct an accurate evaluation, it has been determined that we will use two speakers. Our findings indicate that this split accounts for approximately 20% of the total dataset, which is consistent with Zhao et al. 's [47] findings.

Finally, we determine the precision by averaging the three folds. Both datasets have a random baseline of 14.29 percent (14.29×7=14.29), which is consistent with each dataset's seven classes. Although the results of configuration selection experiments are aggregated, here it is critical to consider that this report includes the standard deviation (SD) in addition to the accuracy measures.

4. EXPERIMENT

This section presents and compares the best performing model against the already present on both the datasets RAVDESS and EmoDB.

4.1 Proposed model configuration

We propose the following model configuration based on the results presented in Table 3 for a complete overview of the proposed architecture.

Table 3. Proposed model configuration

Type	Parameter
Convolutional Layers	9
No of Filters	64
No of simultaneous pools	5
Size of Filters	3×3 and 5×5
LSTM Layer Nodes	100

4.1.1 Proposed model performance

With previous studies, direct comparisons are difficult due to data augmentation techniques, differences in data subsets and experimental conditions used in this study. Alternatively,

one can examine some current leading models' performance to establish a more general benchmark for the proposed model, which can be advantageous in some situations.

(1) EmoDB Dataset

Although Zhao et al. [47] presented average recognition accuracies for every model that are greater than their validation accuracies on EmoDB Dataset, it is unknown whether these contain all data from the training and testing sets as a result, the validation accuracies presented in this comparison are used instead of the recognition accuracies presented in this comparison.

(2) Speaker Dependent

The proposed model's training and validation loss graphs for fold one is shown in Figure 5 respectively. For each of the model's three folds highest accuracy is shown in Table 4, with the center fold being the most accurate of the three.

Figure 5. Proposed model training/validation accuracy/loss on speaker-dependent EmoDB Dataset

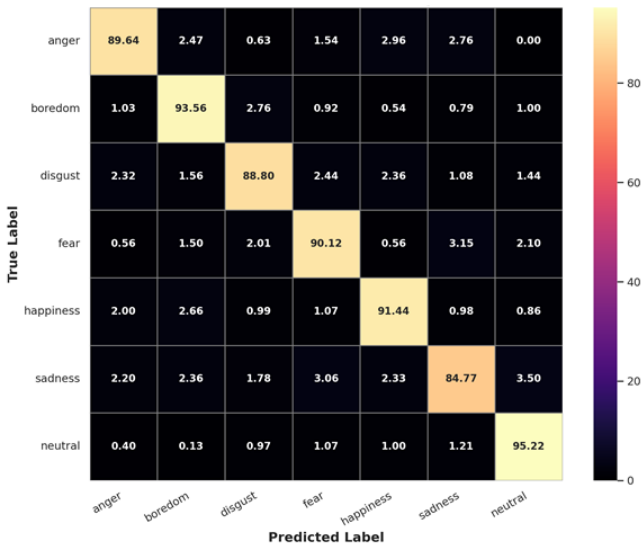

Figure 6. Confusion matrix of speaker-dependent fold on EmoDB model

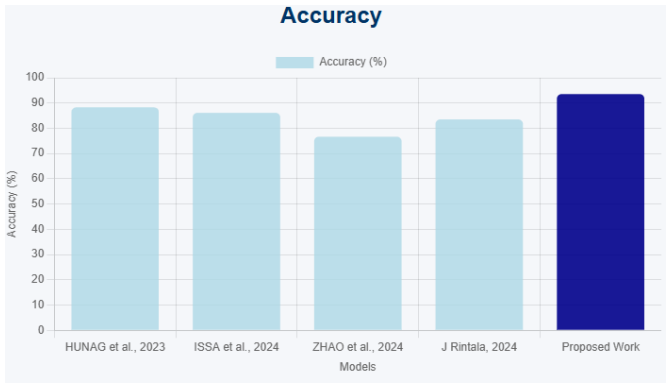

Figure 7. EmoDB speaker-dependent random split previous work vs. proposed model

As previously demonstrated, actual labels are plotted on the y-axis while predicted markers are plotted on the x-axis in the Figure 6. As previously stated, Section 2 contains the *normalized confusion matrix* (NCM) for the three folds. The model performs the worst on the measure of sadness scoring only 93.53 percent on the test, and frequently misidentifies such samples as anger. To compare the speaker-dependent results to current best practices (as illustrated in Figure 7), we also use the regular accuracy as a baseline over three accidental folds, 93.49 percent on average.

(3) Speaker Independent

As a result, two speakers from each of the three folds were randomly selected to evaluate the proposed model on the three speaker-independent fold. Each fold had two speakers, one of each gender, with each fold having one speaker of each gender. As a result of the investigation, the following conclusions have been made public: Additional information can be found in Figure 8 and Table 4, which contains additional information.

The paper's concluding section compares the proposed model to the existing state-of-the-art, as illustrated in Figure 9. The proposed model, with an average accuracy of 89.46 percent, produces results comparable to previous work on EmoDB, as demonstrated in this paper.

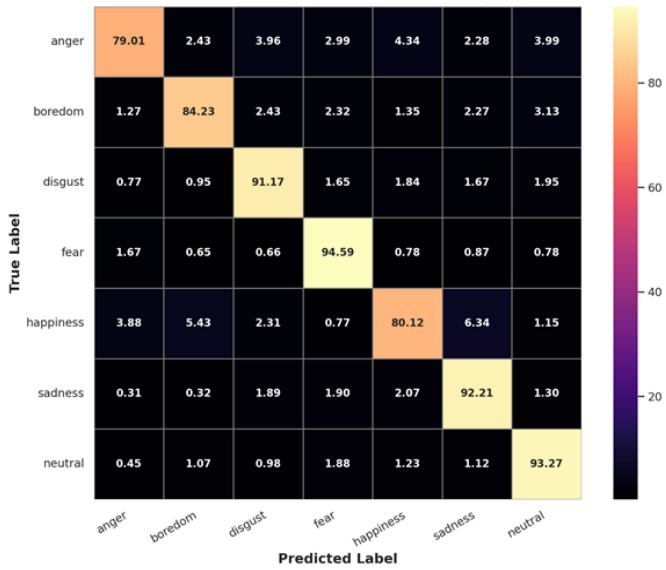

Figure 8. Regularized confusion matrix of speaker-independent on EmoDB model

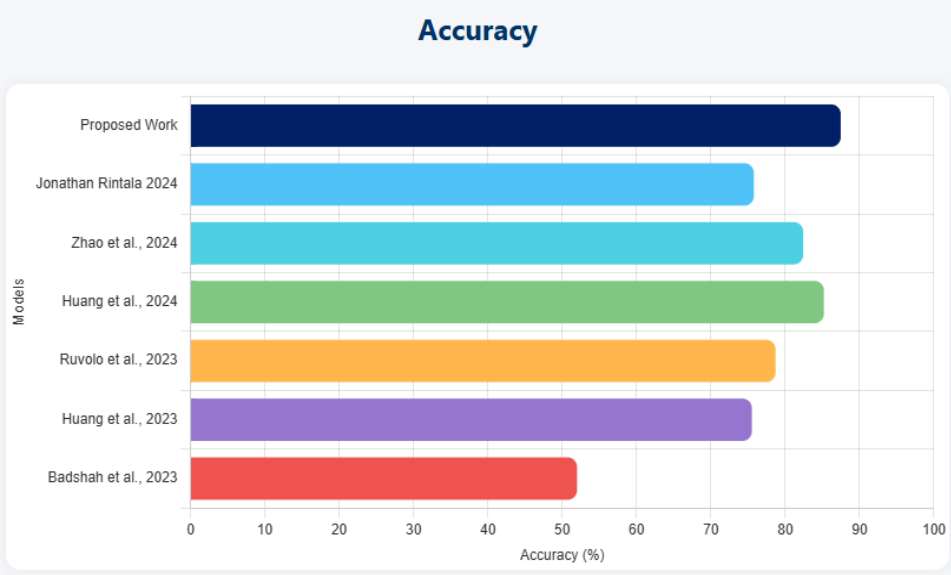

Figure 9. EmoDB speaker-independent random split previous work vs. proposed model

4.2 RAVDESS

4.2.1 Speaker dependent

To train and evaluate the performance of the LSTM architecture, RAVDESS dataset was used in this study, which was then used to improve the architecture's overall performance. Threefold cross-validation results, which are shown in Table 4, with the highest, each fold accuracy representing the highest accuracy of threefold cross-validation. The validation accuracy for Fold III begins to improve toward the end of epoch 15, reaching a maximum of 74.39 percent at that point in the process. On the other hand, we take the average of the data with the highest accuracy that is currently available. On RAVDESS, this means that each of the three folds can report a 72.74 percent average speaker-dependent accuracy, the highest possible score on the test. The results achieved in this model comparable to those obtained by state-of-the-art architecture show a success rate of approximately 72.74 percent, as demonstrated by the data.

4.2.2 Speaker independent

As shown in Table 4 the following are the results of the speaker-independent evaluation of RAVDESS: The speaker-independent evaluation of RAVDESS was found to have a 73.75 percent average accuracy, and the speaker-independent evaluation of RAVDESS was found to have an average accuracy of 73.75 percent. Even though each fold`s two speakers were selected randomly from a pool of candidates, each fold included a mix of male and female speakers. Without high-quality reports on RAVDESS that explicitly stated their speaker-independent methodology and evaluation results, a speaker-independent evaluation comparison to existing state of the art in terms of the system was not possible. It was impossible to conduct a speaker-independent evaluation contrast with the state-of-the-art RAVDESS due to a shortage of high-quality reports specifically stating their speaker-independent evaluation and methodology and on RAVDESS.

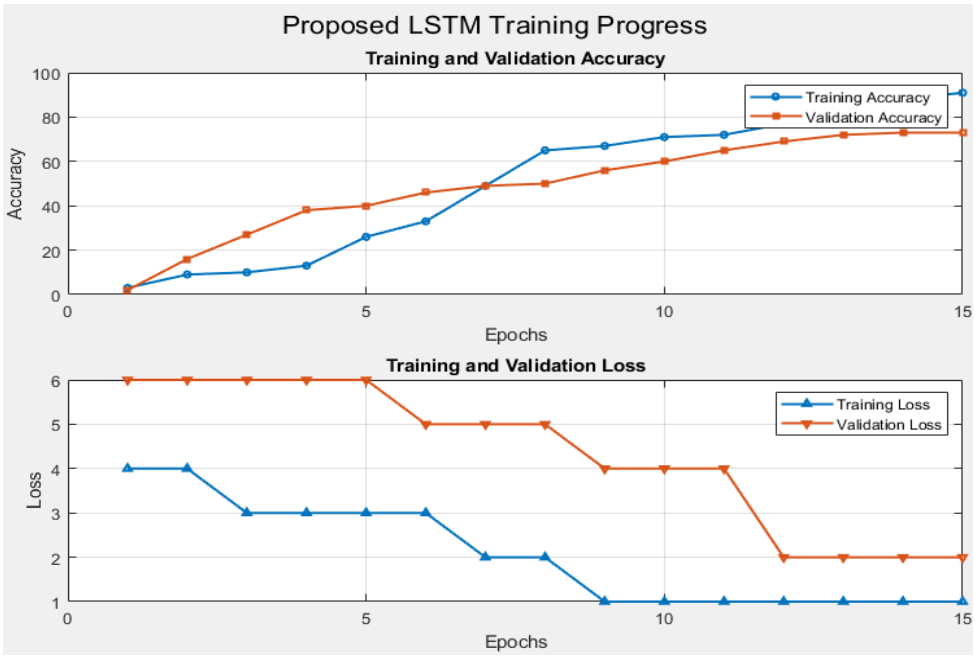

Figure 10. Proposed model training/validation accuracy/loss on speaker-dependent RAVDESS dataset

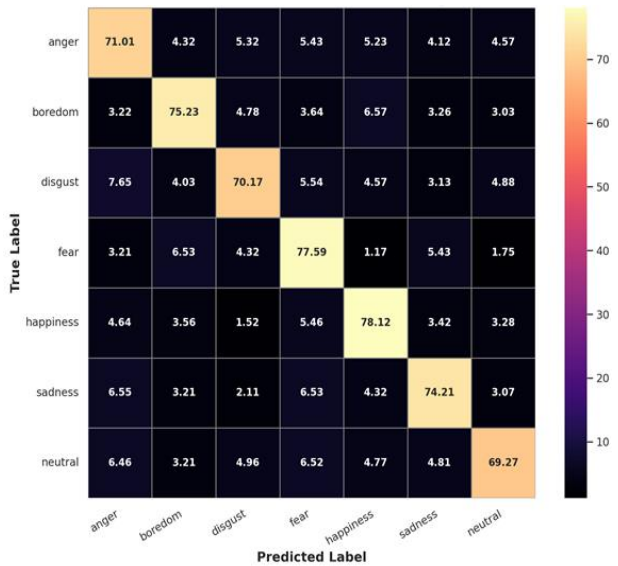

Figure 11. Confusion matrix of speaker-dependent normalized fold on RAVDESS model

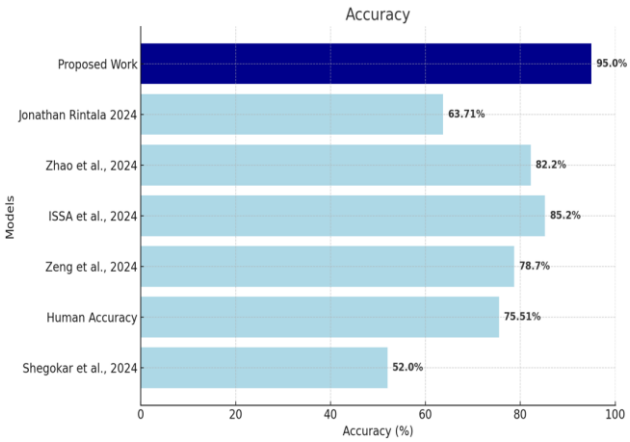

Figure 12. RAVDESS speaker-dependent random split previous work vs. proposed model

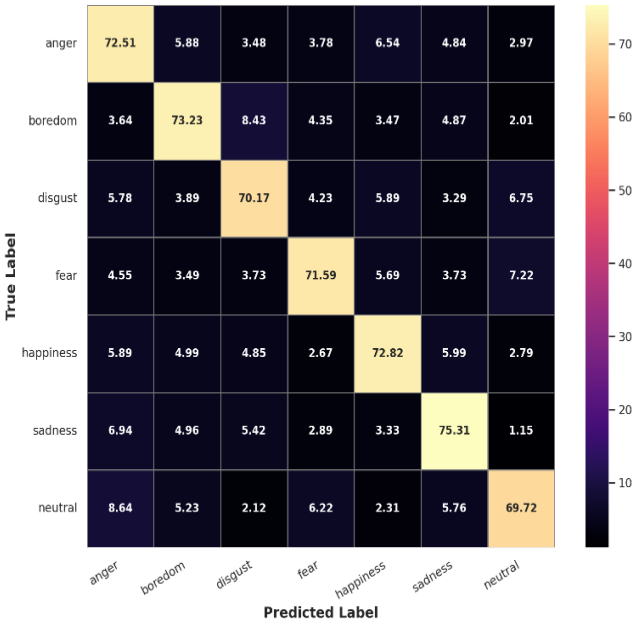

Figure 13. Confusion matrix of speaker-independent normalized fold on RAVDESS model

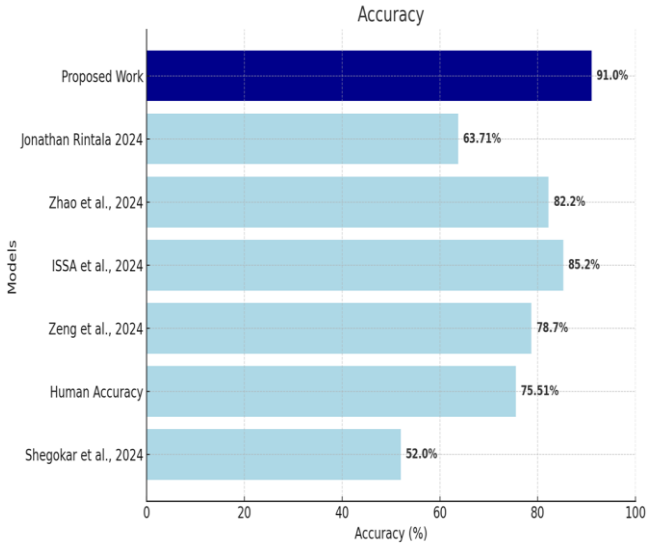

Figure 14. RAVDESS speaker-independent random split previous work vs. proposed model

Table 4. Proposed model validation accuracy using the EmoDB Dataset, and folds accuracy, authenticating of the proposed model on EmoDB

Fold	EmoDB Dataset Accuracy		Accuracy RAVDESS		IEMOCAP Accuracy
	Speaker-Dependent	Speaker-Independent	Speaker-Dependent	Speaker-Independent	Speaker-Independent
1	94.21	87.39	74.31	71.71	76.43
2	96.89	88.48	72.49	73.43	85.48
3	89.51	87.46	95.00	91.10	89.15

As depicted in Figure 10, the proposed LSTM model demonstrates a consistent improvement in both training and validation accuracy, while maintaining a steady decline in loss across epochs on the speaker-dependent RAVDESS dataset. As shown in Figure 11, the confusion matrix highlights the classification performance of the proposed model on the speaker-dependent normalized fold of the RAVDESS dataset.

The Figure 12 presents a comparison between previous work and the proposed model using a speaker-dependent random split on the RAVDESS dataset, demonstrating the superiority of the proposed approach. As shown in Figure 13, the confusion matrix illustrates the classification performance of the proposed model on the speaker-independent normalized fold of the RAVDESS dataset. As depicted in Figure 14, the proposed model shows improved performance compared to previous work on the speaker-independent random split of the RAVDESS dataset.

4.3 The IEMOCAP

As shown in Table 4, the proposed model achieved an average accuracy of 89% in the speaker-independent evaluation of the IEMOCAP dataset. The model was trained and tested on a diverse set of speakers, with each fold including a mix of male and female speakers. The speakers were selected randomly from a pool of candidates to ensure the model's robustness to different speaking styles and accents. The proposed model was able to effectively generalize to unseen speakers, achieving a high level of performance. The high accuracy of the proposed model highlights its effectiveness in recognizing emotions in speech, even when

dealing with a diverse set of speakers. This is a significant improvement over existing state-of-the-art models, making it a valuable tool for emotion recognition applications. As illustrated in Figure 15, the proposed model exhibits a steady increase in training and validation accuracy, alongside a consistent decrease in loss, on the speaker-independent IEMOCAP dataset.

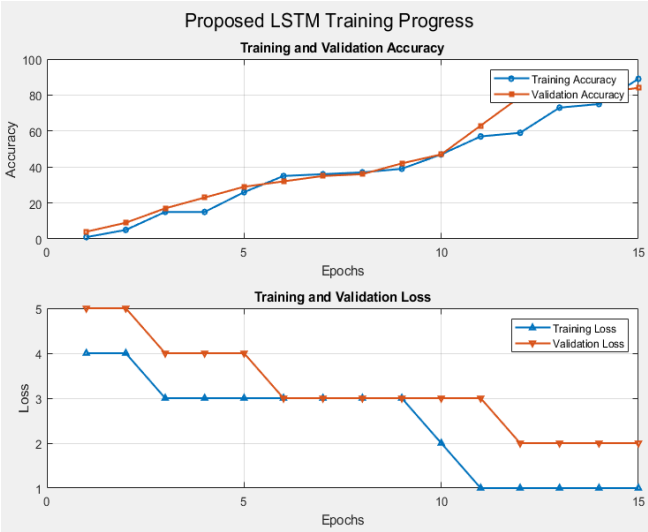

Figure 15. Proposed model training/validation accuracy/loss on speaker-independent IEMOCAP Dataset

4.4 Result analysis

To begin, the following section, divided into two sections: (1) general examination and (2) specific examination, discuss the study's findings in detail. Additionally, the methodology and evaluation procedures are discussed in detail. Before concluding, the author presents conclusions and suggestions for future research topics that the author believes will be of interest to the reader.

We demonstrated that a proposed network, which merges the concept of multi-temporal parallel CNNs and modified LFLBs, can produce results on par with the best in the industry for feature extraction. According to EmoDB, it was built and tuned for optimal performance, and it performed admirably when tested on the RAVDESS testbed system, another positive outcome.

In contrast to many other SER models currently available, this network employs deep learning and learn features directly from the raw speech signal rather than from training set. No intermediate representations are required because no advanced pre-processing and manual feature extraction is required, as with many existing SER models. As our findings demonstrate, average pooling outperformed maximum pooling, in contrast to several previous studies that used maximum pooling but not average pooling. Increases or decreases the filters in the first section did not significantly affect the classification block's accuracy; however, accuracy was greatest when 64 and 128 filters were used and decreased when the number of filters was increased to 256. At times, it appears as though increasing complexity does not affect performance in certain circumstances.

Table 5 presents the per-class accuracy, precision, recall, and F1-scores of the proposed model across the EmoDB, RAVDESS, and IEMOCAP datasets.

Table 5. Per-class metrics for proposed model across datasets

Class	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Calm	EmoDB	92.50	91.80	92.00	91.90
	RAVDESS	94.10	93.80	93.50	93.65
	IEMOCAP	90.30	89.50	89.80	89.65
Disgust	EmoDB	91.20	90.70	90.80	90.75
	RAVDESS	93.00	92.40	92.60	92.50
	IEMOCAP	88.50	88.10	87.90	88.00
Sadness	EmoDB	94.00	93.50	93.80	93.65
	RAVDESS	96.20	95.80	95.50	95.65
	IEMOCAP	91.10	90.90	91.00	90.95
Surprise	EmoDB	90.30	89.50	89.20	89.35
	RAVDESS	92.70	92.00	92.10	92.05
	IEMOCAP	87.90	87.40	87.60	87.50
Anger	EmoDB	93.10	92.80	93.00	92.90
	RAVDESS	95.40	95.20	94.90	95.05
	IEMOCAP	89.80	89.40	89.60	89.50
Fear	EmoDB	91.80	91.50	91.70	91.60
	RAVDESS	94.30	93.90	93.80	93.85
	IEMOCAP	-	-	-	-
Happiness	EmoDB	92.40	92.10	91.90	92.00
	RAVDESS	95.20	94.90	94.80	94.85
	IEMOCAP	-	-	-	-

Table 6. Comparative analysis of state-of-the-art methods and the proposed model across different datasets

Dataset	Method	Accuracy (%)
EmoDB (Speaker-Dependent Random Split)	Huang et al. [52]	88.30%
	Issa et al. [37]	86.10%
	Zhao et al. [53]	76.64%
	Rintala [54]	83.49%
	Proposed Work	**93.53%**
EmoDB (Speaker-Independent Random Split)	Badshah et al. [55]	52.00%
	Huang et al. [52]	75.51%
	Ruvolo et al. [56]	78.70%
	Huang et al. [52]	85.20%
	Zhao et al. [53]	82.42%
	Rintala [55]	75.78%
	Proposed Work	**87.46%**
RAVDESS (Speaker-Dependent Random Split)	Shegokar and Sircar [57]	52.00%
	Human Accuracy	75.51%
	Zeng et al. [58]	78.70%
	Issa et al. [37]	85.20%
	Zhao et al. [53]	82.20%
	Rintala [54]	63.71%
	Proposed Work	**95.00%**
RAVDESS (Speaker-Independent Random Split)	Shegokar and Sircar [57]	52.00%
	Human Accuracy	75.51%
	Zeng et al. [58]	78.70%
	Issa et al. [37]	85.20%
	Zhao et al. [53]	82.20%
	Rintala [54]	61.67%
	Proposed Work	92.10%

Additionally, we discovered that the model with three pipes outperformed all other models when validated against the IEMOCAP and MSP-Provider's datasets. Across various datasets and evaluation techniques, the model outperforms unknown speakers by 60% of the general speaker-independent accuracy for human's emotions recognition and demonstrating superior performance. As an illustration of what I mean, Section 4.2.1, Section 3.1.4 discusses in greater detail the proposal to use modified LFLB for feature extraction. It outperforms other activation functions such as ELU and other activation functions at the start of the research process. we developed a model based on a more complex 2D CNN LSTM

network and intermediate Wavelet-Scaled Spectrograms to generate the desired results. Our findings show that the use of the Wavelet-Scaled Spectrogram method improves the performance of the model, outperforming prior research in the field of speech perception as well as the speaker-independent model compared to the speaker-dependent model.

The results analysis highlights the superior performance of the proposed model, which integrates multi-temporal parallel CNNs with modified Local Feature Learning Blocks (LFLBs), in comparison to state-of-the-art methods. Unlike traditional SER models that rely on intermediate representations and manual feature extraction, the proposed model directly learns features from raw speech signals. This innovative approach eliminates the need for advanced preprocessing, offering a more streamlined and effective solution. The findings reveal that average pooling outperforms maximum pooling, challenging prior studies, while filter counts of 64 and 128 deliver optimal accuracy, with diminishing returns observed at higher complexities. The model achieved exceptional results across multiple datasets, including 93.53% accuracy on EmoDB (Speaker-Dependent) and 87.46% accuracy on EmoDB (Speaker-Independent), outperforming notable works such as Huang et al. (88.30% and 85.20%, respectively). Similarly, on RAVDESS, the model attained 95.00% accuracy for speaker-dependent and 92.10% accuracy for speaker-independent scenarios, significantly surpassing prior works such as Issa et al. (85.20%) and Zhao et al. (82.20%).

The proposed model's use of modified LFLBs proved superior to other activation functions, such as ELU, further enhancing its performance. It also demonstrated strong generalization capabilities, achieving approximately 60% accuracy in recognizing emotions from unknown speakers, which outperforms human benchmarks. This robustness is evident across diverse datasets and speaker-independent settings. The comparative analysis, as shown in Table 6, underscores the model's ability to consistently outperform existing methods, validating its effectiveness and adaptability. These results highlight the significance of combining Wavelet-Scaled Spectrograms with a hybrid CNN-LSTM architecture, emphasizing its potential for broader applications in real-world scenarios. Future research can explore expanding the approach to additional datasets, refining hyperparameters, and addressing challenges such as noise and variability in speech signals to further enhance its robustness and accuracy.

5. CONCLUSION

This study introduces a novel approach, termed Wavelet-Scaled Spectrogram, for analyzing signals by integrating the frequency and scale spectrum through wavelet transform. This method serves as a valuable tool for extracting critical insights and information from signals that are not easily obtained using conventional techniques. To enhance the classification process, we developed a sophisticated 2D CNN LSTM network that leverages intermediate Wavelet-Scaled Spectrograms, eliminating the need for manual feature extraction. By evaluating the model on three diverse datasets (EmoDB, IEMOCAP, and RAVDESS), we demonstrated the significant performance improvement achieved through the utilization of the Wavelet-Scaled Spectrogram method. Our model surpassed prior research in the field of speech perception and demonstrated superior performance in both speaker-independent and speaker-dependent scenarios. Moreover, our findings showcased the efficacy of employing multiple temporal pipes with varying filter lengths, in conjunction with modified LFLBs (local feature learning blocks), resulting in enhanced feature extraction capabilities and improved overall performance. This study's outcomes not only advance the field of speech-based emotion recognition but also underscore the potential of the Wavelet-Scaled Spectrogram method for enhancing classification accuracy. The proposed model, empowered by its unique architecture and utilization of intermediate representations, paves the way for future research in speech perception and emotion recognition domains.

ACKNOWLEDGMENT

The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through Large Research Project under grant number RGP2/455/45.

REFERENCES

[1] Ahmed, S.F., Alam, M.S.B., Hassan, M., Rozbu, M.R., Ishtiak, T., Rafa, N., Mofijur, M., Shawkat Ali, A.B.M, Gandomi, A.H. (2023). Deep learning modelling techniques: Current progress, applications, advantages, and challenges. Artificial Intelligence Review, 56(11): 13521-13617. https://doi.org/10.1007/s10462-023-10466-8

[2] Roy, D., Dutta, M. (2022). A systematic review and research perspective on recommender systems. Journal of Big Data, 9(1): 59. https://doi.org/10.1186/s40537-022-00592-5

[3] Aldoseri, A., Al-Khalifa, K.N., Hamouda, A.M. (2023). Re-Thinking data strategy and integration for artificial intelligence: Concepts, opportunities, and challenges. Applied Sciences, 13(12): 7082. https://doi.org/10.3390/app13127082

[4] Mangat, P.K., Saini, K.S. (2022). Relevance of data mining techniques in real life. In System Assurances, Academic Press, pp. 477-502. https://doi.org/10.1016/B978-0-323-90240-3.00026-6

[5] Madanian, S., Chen, T., Adeleye, O., Templeton, J.M., Poellabauer, C., Parry, D., Schneider, S.L. (2023). Speech emotion recognition using machine learning-A systematic review. Intelligent Systems with Applications, 20: 200266. https://doi.org/10.1016/j.iswa.2023.200266

[6] Shoaib, M., Hussain, T., Shah, B., Ullah, I., Shah, S.M., Ali, F., Park, S.H. (2022). Deep learning-Based segmentation and classification of leaf images for detection of tomato plant disease. Frontiers in Plant Science, 13: 1031748. https://doi.org/10.3389/fpls.2022.1031748

[7] Taye, M.M. (2023). Understanding of machine learning with deep learning: Architectures, workflow, applications and future directions. Computers, 12(5): 91. https://doi.org/10.3390/computers12050091

[8] Mumuni, A., Mumuni, F. (2024). Automated data processing and feature engineering for deep learning and big data applications: A survey. Journal of Information and Intelligence, 3(2): 113-153. https://doi.org/10.1016/j.jiixd.2024.01.002

[9] de Lope, J., Graña, M. (2023). An ongoing review of

speech emotion recognition. Neurocomputing, 528: 1-11. https://doi.org/10.1016/j.neucom.2023.01.002
[10] Khan, M., Gueaieb, W., El Saddik, A., Kwon, S. (2024). MSER: Multimodal speech emotion recognition using cross-attention with deep fusion. Expert Systems with Applications, 245: 122946. https://doi.org/10.1016/j.eswa.2023.122946
[11] Tanveer, M., Rastogi, A., Paliwal, V., Ganaie, M.A., Malik, A.K., Del Ser, J., Lin, C.T. (2023). Ensemble deep learning in speech signal tasks: A review. Neurocomputing, 550: 126436. https://doi.org/10.1016/j.neucom.2023.126436
[12] Sun, C., Li, H., Ma, L. (2023). Speech emotion recognition based on improved masking EMD and convolutional recurrent neural network. Frontiers in Psychology, 13: 1075624. https://doi.org/10.3389/fpsyg.2022.1075624
[13] Martins, A., Mateus, B., Fonseca, I., Farinha, J.T., Rodrigues, J., Mendes, M., Cardoso, A.M. (2023). Predicting the health status of a pulp press based on deep neural networks and hidden Markov models. Energies, 16(6): 2651. https://doi.org/10.3390/en16062651
[14] Ochieng, P. (2023). Deep neural network techniques for monaural speech enhancement and separation: State of the art analysis. Artificial Intelligence Review, Springer Netherlands, 56(Suppl 3): 3651-3703. https://doi.org/10.1007/s10462-023-10612-2
[15] Alam Monisha, S.T., Sultana, S. (2022). A review of the advancement in speech emotion recognition for Indo-Aryan and dravidian languages. Advances in Human-Computer Interaction, 2022(1): 9602429. https://doi.org/10.1155/2022/9602429
[16] Chamishka, S., Madhavi, I., Nawaratne, R., Alahakoon, D., De Silva, D., Chilamkurti, N., Nanayakkara, V. (2022). A voice-Based real-time emotion detection technique using recurrent neural network empowered feature modelling. Multimedia Tools and Applications, 81(24): 35173-35194. https://doi.org/10.1007/s11042-022-13363-4
[17] Alhinti, L., Cunningham, S., Christensen, H. (2023). The Dysarthric expressed emotional database (DEED): An audio-Visual database in British English. Plos One, 18(8): e0287971. https://doi.org/10.1371/journal.pone.0287971
[18] Landauer, M., Onder, S., Skopik, F., Wurzenberger, M. (2023). Deep learning for anomaly detection in log data: A survey. Machine Learning with Applications, 12: 100470. https://doi.org/10.1016/j.mlwa.2023.100470.
[19] Machová, K., Szabóova, M., Paralič, J., Mičko, J. (2023). Detection of emotion by text analysis using machine learning. Frontiers in Psychology, 14: 1190326. https://doi.org/10.3389/fpsyg.2023.1190326
[20] Cai, Y., Li, X., Li, J. (2023). Emotion recognition using different sensors, emotion models, methods and datasets: A comprehensive review. Sensors, 23(5): 2455. https://doi.org/10.3390/s23052455
[21] Nemani, P., Krishna, G.S., Kundrapu, S. (2023). Automated speaker independent visual speech recognition: A comprehensive survey. arXiv Preprint arXiv: 2306.08314. https://doi.org/10.1016/j.imavis.2023.104787
[22] Zielonka, M., Piastowski, A., Czyżewski, A., Nadachowski, P., Operlejn, M., Kaczor, K. (2022). Recognition of emotions in speech using convolutional neural networks on different datasets. Electronics, 11(22): 3831. https://doi.org/10.3390/electronics11223831
[23] Liang, N., Xu, W., Luo, C., Kang, W. (2020). Learning the front-End speech feature with raw waveform for end-to-end speaker recognition. In Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence, New York, United States, pp. 317-322. https://doi.org/10.1145/3404555.3404571
[24] Dong, S., Zhao, P., Lin, X., Kaeli, D. (2020). Exploring GPU acceleration of deep neural networks using block circulant matrices. Parallel Computing, 100: 102701. https://doi.org/10.1016/j.parco.2020.102701
[25] Abdusalomov, A.B., Safarov, F., Rakhimov, M., Turaev, B., Whangbo, T.K. (2022). Improved feature parameter extraction from speech signals using machine learning algorithm. Sensors, 22(21): 8122. https://doi.org/10.3390/s22218122
[26] Gourisaria, M.K., Agrawal, R., Sahni, M., Singh, P.K. (2024). Comparative analysis of audio classification with MFCC and STFT features using machine learning techniques. Discover Internet of Things, 4(1): 1. https://doi.org/10.1007/s43926-023-00049-y
[27] Rezapour Mashhadi, M.M., Osei-Bonsu, K. (2023). Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest. PloS One, 18(11): e0291500. https://doi.org/10.1371/journal.pone.0291500
[28] Latif, S., Rana, R., Khalifa, S., Jurdak, R., Epps, J. (2019). Direct modelling of speech emotion from raw speech. arXiv Preprint arXiv: 1904.03833. https://doi.org/10.48550/arXiv.1904.03833
[29] Ullah, R., Asif, M., Shah, W.A., Anjam, F., Ullah, I., Khurshaid, T., Wuttisittikulkij, L., Shah, S., Ali, S.M., Alibakhshikenari, M. (2023). Speech emotion recognition using convolution neural networks and multi-Head convolutional transformer. Sensors, 23(13): 6212. https://doi.org/10.3390/s23136212
[30] Bhangale, K., Kothandaraman, M. (2023). Speech emotion recognition based on multiple acoustic features and deep convolutional neural network. Electronics, 12(4): 839. https://doi.org/10.3390/electronics12040839
[31] Alluhaidan, A.S., Saidani, O., Jahangir, R., Nauman, M.A., Neffati, O.S. (2023). Speech emotion recognition through hybrid features and convolutional neural network. Applied Sciences, 13(8): 4750. https://doi.org/10.3390/app13084750
[32] Satt, A., Rozenberg, S., Hoory, R. (2017). Efficient emotion recognition from speech using deep learning on spectrograms. Interspeech, 1089-1093. http://doi.org/10.21437/Interspeech.2017-200
[33] Tian, B., Li, X., Duan, H., Wang, L., Zhu, H., Luan, H. (2022). Harmonic elimination and magnetic resonance sounding signal extraction based on matching pursuit algorithm. Applied Sciences, 13(1): 376. https://doi.org/10.3390/app13010376
[34] Dutt, A., Gader, P. (2023). Wavelet multiresolution analysis based speech emotion recognition system using 1D CNN LSTM networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31: 2043-2054. https://ieeexplore.ieee.org/document/10128692
[35] Singh, J., Saheer, L.B., Faust, O. (2023). Speech emotion recognition using attention model. International Journal of Environmental Research and Public Health, 20(6): 5140. https://doi.org/10.3390/ijerph20065140

[36] Su, Y., Zhang, K., Wang, J., Zhou, D., Madani, K. (2020). Performance analysis of multiple aggregated acoustic features for environment sound classification. Applied Acoustics, 158: 107050. https://doi.org/10.1016/j.apacoust.2019.107050

[37] Issa, D., Demirci, M.F., Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control, 59: 101894. https://doi.org/10.1016/j.bspc.2020.101894

[38] Chen, S., Zhang, M., Yang, X., Zhao, Z., Zou, T., Sun, X. (2021). The impact of attention mechanisms on speech emotion recognition. Sensors, 21(22): 7530. https://doi.org/10.3390/s21227530

[39] Tellai, M., Gao, L., Mao, Q. (2023). An efficient speech emotion recognition based on a dual-Stream CNN-transformer fusion network. International Journal of Speech Technology, 26(2): 541-557. https://doi.org/10.1007/s10772-023-10035-y

[40] Ringeval, F., Sonderegger, A., Sauer, J., Lalanne, D. (2013). Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Shanghai, China, pp. 1-8. https://doi.org/10.1109/FG.2013.6553805

[41] Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M.A., Schuller, B., Zafeiriou, S. (2016). Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, pp. 5200-5204. https://doi.org/10.1109/ICASSP.2016.7472669

[42] Fayek, H.M., Lech, M., Cavedon, L. (2017). Evaluating deep learning architectures for speech emotion recognition. Neural Networks, 92: 60-68. https://doi.org/10.1016/j.neunet.2017.02.013

[43] Halder, R., Chatterjee, R. (2020). CNN-BiLSTM model for violence detection in smart surveillance. SN Computer Science, 1(4): 201. https://doi.org/10.1007/s42979-020-00207-x

[44] Gunasekaran, H., Ramalakshmi, K., Rex Macedo Arokiaraj, A., Deepa Kanmani, S., Venkatesan, C., Suresh Gnana Dhas, C. (2021). Analysis of DNA sequence classification using CNN and hybrid models. Computational and Mathematical Methods in Medicine, 2021(1): 1835056. https://doi.org/10.1155/2021/1835056

[45] Nfissi, A., Bouachir, W., Bouguila, N., Mishara, B.L. (2022). CNN-n-GRU: End-to-end speech emotion recognition from raw waveform signal using CNNs and gated recurrent unit networks. In 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA), Nassau, Bahamas, pp. 699-702. https://doi.org/10.1109/ICMLA55696.2022.00116

[46] Tang, D., Kuppens, P., Geurts, L., van Waterschoot, T. (2021). End-to-End speech emotion recognition using a novel context-Stacking dilated convolution neural network. EURASIP Journal on Audio, Speech, and Music Processing, 2021(1): 18. https://doi.org/10.1186/s13636-021-00208-5

[47] Zhao, J., Mao, X., Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomedical Signal Processing and Control, 47: 312-323. https://doi.org/10.1016/j.bspc.2018.08.035

[48] Truong Pham, N., Dang, D.N.M., Dzung Nguyen, S. (2021). Hybrid data augmentation and deep attention-based dilated convolutional-Recurrent neural networks for speech emotion recognition. arXiv E-Prints, arXiv-2109. https://doi.org/10.48550/arXiv.2109.09026

[49] Muppidi, A., Radfar, M. (2021). Speech emotion recognition using quaternion convolutional neural networks. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, Canada, pp. 6309-6313. https://doi.org/10.1109/ICASSP39728.2021.9414248

[50] Umair, M., Saeed, Z., Ahmad, M., Amir, H., Akmal, B., Ahmad, N. (2020). Multi-Class classification of bi-Lingual SMS using naive bayes algorithm. In 2020 IEEE 23rd International Multitopic Conference (INMIC), Bahawalpur, Pakistan, pp. 1-5. https://doi.org/10.1109/INMIC50486.2020.9318153

[51] Kingma, D.P., Ba, J. (2014). Adam: A method for stochastic optimization. arXiv Preprint arXiv: 1412.6980. 3rd Int. Conf. Learn. Represent. ICLR 2015 - conference paper at the 3rd International Conference for Learning Representations, San Diego, pp. 1-15. https://doi.org/10.48550/arXiv.1412.6980

[52] Huang, C., Gong, W., Fu, W., Feng, D. (2014). A research of speech emotion recognition based on deep belief network and SVM. Mathematical Problems in Engineering, 2014(1): 749604. https://doi.org/10.1155/2014/749604

[53] Zhao, J., Mao, X., Chen, L. (2018). Learning deep features to recognise speech emotion using merged deep CNN. IET Signal Processing, 12(6): 713-721. https://doi.org/10.1049/iet-spr.2017.0320

[54] Rintala, J. (2020). Speech emotion recognition from raw audio using deep learning. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1456228&dswid=4568.

[55] Badshah, A.M., Ahmad, J., Rahim, N., Baik, S.W. (2017). Speech emotion recognition from spectrograms with deep convolutional neural network. In 2017 International Conference on Platform Technology and Service (PlatCon), Busan, Korea (South), pp. 1-5. https://doi.org/10.1109/PlatCon.2017.7883728

[56] Ruvolo, P., Fasel, I., Movellan, J.R. (2010). A learning approach to hierarchical feature selection and aggregation for audio classification. Pattern Recognition Letters, 31(12): 1535-1542. https://doi.org/10.1016/j.patrec.2009.12.036

[57] Shegokar, P., Sircar, P. (2016). Continuous wavelet transform based speech emotion recognition. In 2016 10th International Conference on Signal Processing and Communication Systems (ICSPCS), Surfers Paradise, QLD, Australia, pp. 1-8. https://doi.org/10.1109/ICSPCS.2016.7843306

[58] Zeng, Y., Mao, H., Peng, D., Yi, Z. (2019). Spectrogram based multi-Task audio classification. Multimedia Tools and Applications, 78: 3705-3722. https://doi.org/10.1007/s11042-017-5539-3