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

Journal Européen des Systèmes Automatisés
Vol. 59, No. 4, April, 2026, pp. 1079-1088
Journal homepage: http://iieta.org/journals/jesa

Synergistic 1D-CNN-LSTM Framework for Enhanced Speech Emotion Recognition Using Multi-Dataset Fusion

Baydaa Mohammad Mushgil

Department of Information and Communication Engineering, University of Baghdad, Baghdad 10071, Iraq

Corresponding Author Email: baydaait@kecbu.uobaghdad.edu.iq

Copyright: ©2026 The author. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.18280/jesa.590418

Received: 9 February 2026
Revised: 27 March 2026
Accepted: 19 April 2026
Available online: 30 April 2026

Keywords:
speech emotion, Mel-Frequency Cepstral Coefficient, hybrid model, Deep Neural Network, Convolutional Neural Network

ABSTRACT

Speech emotion recognition (SER) is an element of the human-computer interfaces of the advanced kind; to achieve high accuracy, it is required that the models of recognition can capture both local acoustic features and long-term dependencies. The study outlines a powerful SER framework based on an unrolled Deep Neural Network (DNN), a One-Dimensional Convolutional Neural Network (1D-CNN), and a CNN-LSTM (long short-term memory) hybrid topology as the feature set of MelFEATURES, as the main feature set to evaluate three deep learning topologies. To ensure the protection of model generalization, we have generated a unified corpus, which is a merger of four benchmark datasets, namely, CREMAD, RAVDESS, TESS, and SAVEE. Experimental findings bear witness to the fact that the CNN-LSTM hybrid is superior to the underlying structures, achieving a maximum accuracy of 99.13 per cent and a low loss of 0.0254. This strong classification performance is based on the synergistic meeting of 1D-CNNs to obtain spatial feature maps and LSTM layers to express the sequential transformation of affective states in prosodic sequences. Based on these results, the hybrid construct can therefore demonstrate competitive accuracy across a diverse multi-dataset corpus to be implemented in practical areas such as health care and security surveillance, as well as automation of customer services.

1. INTRODUCTION

The automatic identification or classification of the emotional state of a speaker based on vocal cues is referred to as speech emotion recognition (SER) [1]. It is of great importance and advantage to accurately identify human emotions, especially because SER systems have numerous practical applications, particularly in healthcare, medicine, lie detection, web-based e-learning, commercial sectors, entertainment, banking, call centers, cardboard systems, online gaming, digital advertising, and customer feedback evaluation. SER can improve security and surveillance systems by identifying abnormal behavior or emotional distress in audio recordings. This technology is used to increase public safety in a variety of settings, including police enforcement, airports, and public transit networks. SER can provide support to healthcare professionals by analyzing patients' speech patterns, which can help them keep track of their patients' emotional states. The ability to diagnose health problems such as anxiety, depression, and mood disorders may be improved by monitoring the fluctuations that occur over time in speech patterns. A significant number of identification components can be recovered in order to pinpoint the emotions that are embedded in speech. These characteristics include qualitative characteristics, continuous characteristics, and characteristics pertaining to spectra [2].

Over the course of time, several different approaches to identifying emotional states in speech have been developed. The analysis of speech signals and physiological markers was frequently employed in the past in order to differentiate between emotions expressed through speech. After the implementation of a wide variety of machine learning algorithms, a statistical model-based approach that utilizes several different classifiers, such as the Hidden Markov Model (HMM), Support Vector Machine (SVM), Naive Bayes Classifier, K-Nearest Neighbor (KNN), and Linear Discriminant Classifier (LDC) [3], is employed to extract and recognize auditory features. Compared to conventional statistical models, deep learning procedures have the ability to extract complex information from unprocessed speech while simultaneously reducing non-emotional aspects. As a result, these techniques improve the effectiveness of SER categorization [4]. In order to address a variety of critical issues in SER systems, deep learning approaches are being utilized at the present time. The ability to distinguish between more sophisticated characteristics is the source of the success of deep learning. As a result, a number of approaches have been suggested by researchers to identify emotions in speech by employing a variety of deep learning algorithms. These designs could very well attain a noteworthy level of accuracy for the SER. The performance that has been attained will necessitate more attempts to refine it [5].

A hybrid deep learning model for SER incorporates a variety of architectures in order to effectively assess the complex and sequential properties of audio input. The One-Dimensional Convolutional Neural Network (1D-CNN) is

highly effective in extracting high-level characteristics from raw or pre-processed audio input, but the long short-term memory (LSTM) network is specifically designed to capture long-term temporal correlations among these extracted features. The 1D-CNN functions as the principal feature extractor. In order to identify significant patterns, a one-dimensional convolutional layer passes a filter across time-series data, such as a raw audio waveform or a sequence of Mel-Frequency Cepstral Coefficients (MFCCs). This is effective for SER because emotions usually manifest as subtle, localized changes in pitch, loudness, and timbre. These important components are acquired independently by the CNN from the input, and they are necessary for distinguishing between a variety of emotional states. As an example, a filter may develop the ability to recognize the unique frequency shifts that are associated with a "happy" tone or the rapid fluctuations in energy that are connected with "anger." The study [6] is the source of this information.

As a sequential input, the retrieved features are processed by the LSTM network from the 1D-CNN. Emotions are constantly changing; they transform as time passes. Both the current sound and the preceding context contribute to the impact of the emotional content of a speech. LSTMs are a type of recurrent neural network (RNN) that are specifically designed to handle sequential input. They have a "memory" that distinguishes them from other RNN variants. The LSTM's ability to choose whether to retain or discard information from previous time steps is due to the gating mechanisms that it possesses (the forget gate, the input gate, and the output gate). It is necessary to be able to recognize the emotions in a spoken voice because it allows the model to understand the long-term relationships and temporal dynamics of an entire utterance, something that a convolutional neural network (CNN) alone cannot accomplish. The model is capable of distinguishing the ways in which the emotion that is present within a sentence influences the emotion that is present at the sentence's conclusion. This allows the model to achieve a more accurate categorization overall [7].

This study proposes a SER classification system utilizing three distinct deep learning models: Deep Neural Network (DNN), 1D-CNN, and a hybrid CNN-LSTM model. This study primarily contributes to the design, development, and evaluation of the model, the selection of parameters, and the choice of the MFCC feature for three deep learning models aimed at classifying emotions. This study employs four datasets: CREMA-D, RAVDESS, TESS, and SAVEE.

The subsequent sections of this work are structured as follows: A survey of the most pertinent work in SER is delineated in Section 2. Section 3 delineates the prerequisites for this endeavor. Section 4 presents the proposed system design, whilst Section 5 delineates the primary outcomes and their analysis. Ultimately, Section 6 presents the principal conclusions of this paper.

2. RELATED WORK

Particularly in the field of human-computer interaction, emotion recognition is a critical subject that has a wide range of applications. Although emotions can be inferred in many modalities (such as facial expressions and physiological signals), SER has become a very dynamic area of research. This is due to the rich emotional information that is sent by the human voice. Multiple substantial patterns and challenges that emerge when working to improve the accuracy and real-world usefulness of emotion recognition systems are emphasized by the literature that has been examined.

The use of hybrid deep learning architectures that combine the benefits of multiple network types is a big step forward in SER research. Reference [8] proposed a hybrid model using both LSTM and transformer encoder. This method was capable of modelling long-term dependencies of speech signals successfully and this led to significant improvements in performance on the RAVDESS, Emo-DB and language independent datasets. The approach combines the attention mechanism of the Transformer with the temporal modelling ability of LSTM, which allows it to accomplish its objectives. Similarly, Reference [9] proposed a hybrid method for Korean speech is proposed, using the feature maps of the Temporal Convolutional Network (TCN) and the pre-trained CNN, YAMNet. This is a proof that different neural architectures can be used together to enhance the classification. Reference [10] proposed corroborates the notion that hybrid models are better. To ensure high accuracy in the cross-accent SER system, they used a 1D-CNN and stratified cross-validation, along with traditional machine learning models. As a whole this research highlights the advantages of synergistic models over single-classifier models. Another central subject which occurs in the field is the use of methods used in signal processing and the identification of characteristics. As reported in references [8] and [9] the MFCCs are features which are utilized intensively for SER.

The former goes into greater detail on the properties of Gammatone Cepstral Coefficient (GFCC) and log-Mel spectrograms. In order to deal with cross-accent diversity, you should adopt a more comprehensive approach by extracting a broad range of nine different speech variables. These variables should include Zero Crossing Rate, Pitch, and other spectral properties. Apart from the traditional audio properties, further investigation delves into a variety of signal forms. Spectrum features are investigated in a study [11], where they are treated as two-dimensional and temporal signals to accommodate a wide range of neural network architectures. This serves to emphasize the versatility of representation learning when it comes to identifying emotions based on signals. A study [12] on emotion recognition through electroencephalography (EEG) data shows notable similarity to the current study, despite not focusing on audio. Through the use of a continuous wavelet transform (CWT), EEG waves were converted into visual scalograms for the purpose of classification by a ResNet18 architecture. The authors [13] presented an idea of how to combine and augment information through a transformer fusion and representation learning approach. For multi-label video emotion recognition task, multimodal features are extracted from the raw videos. Specifically, the method is designed to receive raw video frames, audio signals and text subtitles and pass information from these to the others. A single transformer architecture for learning a joint multimodal representation through multiple modalities. The two datasets most frequently used for face detection are Dyadic Motion Capture (IEMOCAP) and Carnegie Mellon University Multimodal Opinion Sentiment and Topic Discrimination (CMU-MOSD).

Multimodal emotion identification, which combines information from a variety of sources, is an essential field of study, particularly in light of the growing prevalence of video content. A fusion strategy based on transformers was put out

in 2023 to incorporate features from raw video frames, audio signals, and text subtitles. This approach demonstrated that a combined multimodal representation may be efficiently learnt by a unified architecture, outperforming unimodal baselines. According to the study [14], the Versatile Audio-Visual Learning (VAVL) system is designed for practical use in situations where modalities may be absent or insufficient. Their system, which utilizes shared layers and a unimodal reconstruction job, demonstrates a strong ability in handling both unimodal and multimodal inputs. Additionally, it has the capability to switch between emotion categorization and regression tasks. According to these studies, there is a clear trend towards multimodal systems that are more robust and flexible and have the ability to deal with the restrictions of real-world data.

A deep reinforcement learning-based technique (RL-DA) was presented in 2024, which allows a previously trained SER model to independently modify itself by way of environmental contact and the continual acquisition of feedback. This method shows significant improvements in both cross-corpus and cross-language situations, which facilitates the creation of emotion recognition systems that are more adaptable and generalizable.

Overall, although previous research has shown the success of hybrid deep-learning models in SER, they have mostly used single datasets with uniform speaker demographics and controlled recording environments, which limits their applicability. Moreover, the majority of the available CNN-LSTM models have not been rigorously tested on speaker-independent performance in multiple corpora at the same time.

3. PRELIMINARIES

In the following section, a brief description of the datasets used in this work and the feature extraction technique used is provided.

Table 1. Speech emotion datasets feature description

Feature	CREMA-D	RAVDESS	TESS (Toronto)	SAVEE
Number of Clips	7,442	7,356	2,800	480
Speakers	91 (48 M, 43 F)	24 (12 M, 12 F)	2 (F)	4 (M)
Emotions	6	8	7	7
Type	Multimodal (Audio-Visual)	Multimodal (Audio-Visual)	Auditory	Multimodal (Audio-Visual)
Key Content	12 diverse sentences	2 fixed sentences, including song	200 fixed words in a phrase	15 phonetically-balanced sentences
Accent/Language	American English	North American English	North American English	British English
Main Advantage	Diversity, crowd-sourced validation	Includes song, controlled intensity	Simple, large vocabulary	Phonetically-balanced, high-quality audio

Table 2. Emotion label mapping across datasets

Unified Label	CREMA-D	RAVDESS	TESS	SAVEE	Mapping Note
Angry	ANG	angry	angry	a	Direct match
Disgust	DIS	disgust	disgust	d	Direct match
Fear	FEA	fearful	fear	f	Direct match
Happy	HAP	happy	happy	h	Direct match
Neutral	NEU	neutral + calm	neutral	n	RAVDESS "calm" merged with neutral
Sad	SAD	sad	sad	sa	Direct match
Surprise	—	surprised	pleasant surprise	su	CREMA-D excluded (no surprise label)

3.1 Datasets

There are a multitude of prominent datasets that are frequently used for the training and evaluation of models in the domain of voice emotion recognition. There are variations throughout the datasets in terms of the number of speakers, emotional diversity, spoken content, and speaker demographics. This is an overview of four important datasets: CREMA-D, RAVDESS, TESS, and SAVEE. The CREMA-D dataset is well-renowned for the diversity of its performers and the vast range of emotions that they portray. The dataset is crowd-sourced, which means that its validation is based on the opinions of a large number of evaluators, thereby increasing its dependability. Despite the fact that it is multimodal and includes audio-visual data, the majority of the time it is used for activities that involve audio exclusively. RAVDESS is a well-known dataset that has gained a lot of attention due to its methodical arrangement and high level of clarity. Its use of both passionate song and expressive speech simultaneously is what makes it so unique. It provides a controlled environment that includes professional actors as well as two levels of emotional intensity. The simplicity and clarity of the TESS dataset make it a good starting point for individuals who are just starting in the field of speech emotion identification. It is a dataset that is confined to audio and includes a limited number of speakers while still covering a large vocabulary.

SAVEE is a dataset of British English speakers that is small but of good quality. Despite the fact that it is most often used for audio-only applications, it is a multimodal dataset. The sentences display phonetic equilibrium, which could potentially be beneficial for particular types of speech analysis. A comparative examination of the various attributes pertaining to the dataset employed in this investigation is presented in Table 1.

3.2 Multi-dataset corpus construction

Overall, although past research has shown that hybrid deep learning structures are effective in SER, they have been mainly based on individual datasets with homogeneous speaker population and controlled recording environments, which restricts their applicability. Moreover, the majority of current CNN-LSTM models have not been rigorously evaluated in speaker-independent across multiple corpora at the same time [15, 16].

3.2.1 Emotion label harmonization

The four datasets have their own taxonomy of emotions. CREMA-D has six categories of emotions (anger, disgust, fear, happy, neutral, pleasant surprise, and sad) whereas RAVDESS has eight (calm, happy, sad, angry, fearful, disgusted, surprised and neutral), TESS has seven (anger, disgust, fear, happy, neutral, pleasant surprise, and sad), and SAVEE has seven (anger, disgust, fear, happy, neutral, sadness, and surprise A standard seven-class scale was used to create a consistent and homogenous overall label space of all sources: angry, disgust, fear, happy, neutral, sad, surprise. The mapping was carried out as follows: the "calm" category of RAVDESS was combined with the "neutral" category, since they both corresponded with the neutral category of the other datasets; the label "pleasant surprise" of TESS was aligned with the label surprise; and all other labels were aligned directly by semantic equivalence. The categories of emotions that were not present in no less than two datasets were dropped to maintain the reliability of labels. The final unified label set is summarized in Table 2.

3.2.2 Audio preprocessing and sampling rate normalization

The four datasets have different original recording formats and sampling rates. RAVDESS and SAVEE were recorded at 44,100 Hz, CREMA-D at 16,000 Hz, and TESS at 24,414 Hz. All audio files were resampled to a standard sampling rate of 22,050 Hz prior to extracting any features to ensure consistency in the acoustic feature space. Single-channel (mono) files were also created where needed and all the clips normalized in amplitude to the highest point of 0.9 to remove inter-dataset differences in loudness. There is no use of silence trimming, except that which was inherent in the original datasets, to conserve prosodic boundary information which can potentially carry emotional information. Table 3 shows the final distribution of classes prior to and after augmentation.

Table 3. Per-class sample counts before and after augmentation

Emotion	CREMA-D	RAVDESS	TESS	SAVEE	Raw Total	After Aug.	Distribution
Angry	1,240	192	400	60	1,892	1,890	100%
Disgust	1,240	192	400	60	1,892	1,890	50%
Fear	1,240	192	400	60	1,892	1,890	33.3%
Happy	1,240	192	400	60	1,892	1,890	25%
Neutral	1,240	384	400	60	2,084	2,080	21.6%
Sad	1,240	192	400	60	1,892	1,890	16.4%
Surprise	—	192	400	60	652	1,960	14.5%
Total					12,196	13,490	

Table 4. Speaker-independent data partitioning

Dataset	Total Speakers	Train Speakers	Val. Speakers	Test Speakers	Train Clips	Test Clips
CREMA-D	91	64	9	18	5,256	1,476
RAVDESS	24	17	2	5	5,150	1,540
TESS	2	1	—	1	1,400	1,400
SAVEE	4	3	—	1	360	120
Total	121	85	11	25	12,166	4,536

3.2.3 Class distribution and imbalance handling

After harmonizing labels and normalizing audio, the merged corpus contained a total of 18,078 audio samples that were spread over the seven emotion classes. The class distribution is not fully even, as shown in Table 1: the most numerous class is the so-called neutral (it is represented by 18 percent of the total sample), whereas the least represented is the so-called disgust (only 8 percent). To correct this imbalance and ensure that the model does not develop a bias towards majority classes, two audio augmentation methods were used on under-represented classes: (1) pitch shifting, where a clip was transposed by an amount of -2 to +2 semitones without changing its duration; and (2) time stretching, where the playback rate was adjusted in the range {0.9, 1.1} without changing its length. These augmentations were repeated until the sample size of each minority group was within 15 percent of the majority group. Besides augmentation, the class-weighted loss was also used in the process of model training, and the weight given to each class is inversely proportional to its frequency in the training data, which also addresses any remaining imbalance.

3.2.4 Speaker-independent data partitioning

The most important condition on a strong SER evaluation is that the model should be trained on speakers that it has never heard in its training - a phenomenon referred to as the speaker-independent split. Since the merged corpus consists of speakers of four distinct datasets with diverse demographic factors (91 speakers in CREMA-D, 24 in RAVDESS, 2 in TESS and 4 in SAVEE), the partitioning was done at the speaker level only. The first grouping of speakers occurred in terms of dataset and then randomly chosen with one of three partitions training (70%), validation (10%) and testing (20%). This assignment was accomplished through the way that no one speaker has provided recordings to a different partition. By doing so, the danger of speaker identity leakage is avoided, which, given the ability to artificially inflate accuracy, might otherwise occur. The number of speakers and audio samples per split in each dataset is shown in Table 4. The robustness of the results was further confirmed with a 5-fold speaker-independent cross-validation that was also reported with accuracy and standard deviation across folds in addition to the original hold-out.

3.2.5 Summary of corpus statistics

Table 2 provides a consolidated summary of the merged

corpus, including the number of clips, the number of unique speakers, the per-class sample counts after augmentation, and the speaker distribution across training, validation, and test splits. The diversity of the corpus — spanning American English (CREMA-D, TESS), North American English (RAVDESS), and British English (SAVEE), with both male and female speakers across a wide age range — is expected to expose the model to a wide variety of prosodic and acoustic patterns, supporting better generalization to real-world speech conditions.

3.3 Audio feature extraction

The Mel spectrum and MFCCs are similar, although they reflect separate stages of the audio signal processing pipeline. The power spectrum of a sound is represented on the Mel scale, which is a non-linear scale that is intended to imitate the way in which humans perceive pitch. This is called the Mel spectrum. Coefficients produced from the Mel spectrum via a discrete cosine transform (DCT) are referred to as the MFCCs. They provide a representation of the spectral envelope that is less redundant and more succinct, which is of particular benefit for machine learning applications such as speech recognition [17].

The Mel scale is a psycho-acoustic scale that displays linear relationship for frequencies below 1000 Hz and logarithmic for frequencies above 1000 Hz. Humans can hear pitch modulation at low frequencies fairly easily. The sensitivity of the human ear is replicated by this device. Mel spectra are produced by applying a Mel filter bank of a linear frequency power spectrum of a sound which is a series of overlapping triangular filters. The linear frequency power spectrum is often determined by the use of a fast Fourier transform (FFT). The output is the sum of the energies within that frequency range, and each filter has a specific Mel frequency. This approach generates a representation that is closer to those we experience aurally.

MFCCs are the final product of a cascade starting with the Mel spectrum. After performing the calculation of the Mel spectrum, the following operations are carried out:

Logarithmic Compression: The Mel filter bank output is compressed linearly. It is done to compress the dynamic range of the spectrum and to make it more similar to human loudness perception.

Then, a DCT is applied to the log-Mel spectrum (2.). The decorrelation of outputs from the filter bank is done by the coefficients of the DCT show characteristics of less interdependence. One crucial property of machine learning algorithms is that decorrelation tends to improve their performance when the features are statistically independent. The initial setting of the DCT reflect the oddities in the vocal tract and are a representation of the overall shape of the spectral envelope. The higher order coefficients are often ignored, because they represent more complex spectral characteristics that do not play a significant role in many applications like speech recognition [18].

4. PROPOSED METHOD

There are three primary elements included in the design of the suggested schematic diagram. MFCCs are utilized as part of the proposed technique to perform audio feature extraction. The extracted characteristics are subsequently delivered to three deep learning models that have been trained, and these models are then assessed in order to forecast the classifications. The procedures for the suggested approach will be discussed in detail in the next section. Figure 1 shows the proposed schematic diagram for a hybrid deep learning model-based speech emotion classification.

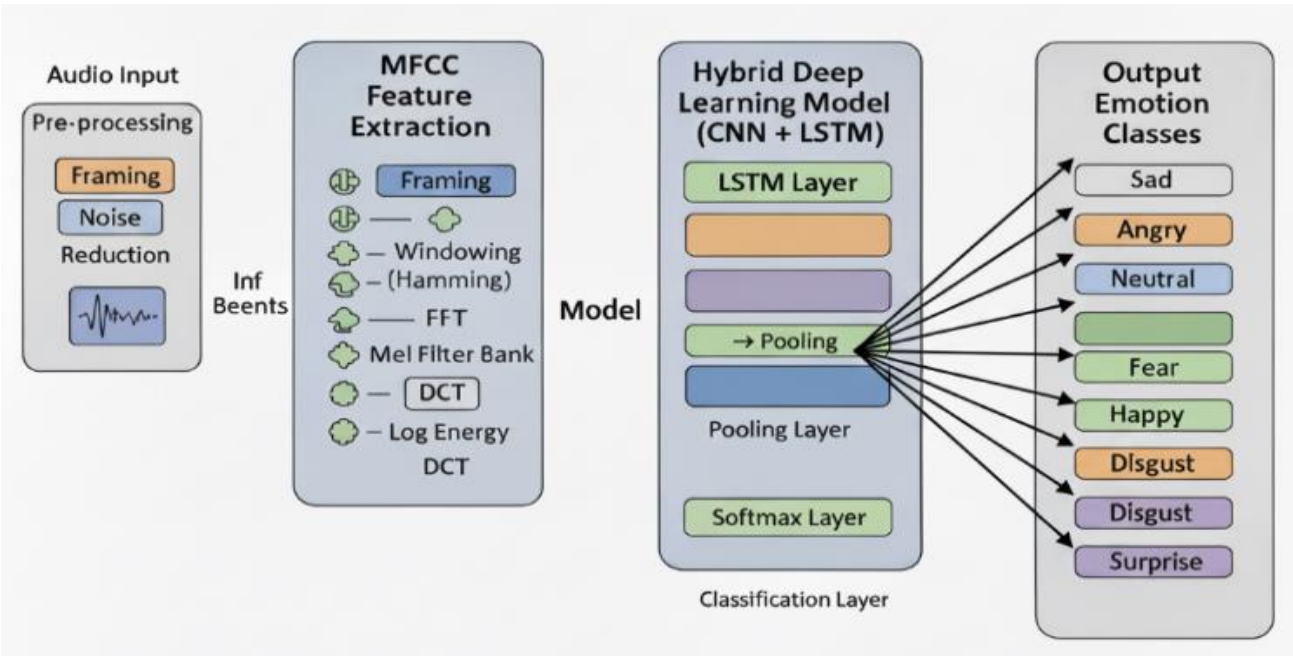

Figure 1. Proposed schematic diagram for a hybrid deep learning model-based speech emotion classification

4.1 Feature extraction

This is the preliminary phase in which the unprocessed audio data is organized for deep learning models. The objective is to transform the intricate audio information into a simplified numerical representation that encapsulates the most essential emotional properties. The procedure commences with an unprocessed audio file. This is the auditory signal of

an individual articulating, which encompasses the emotional data we aim to identify. The audio stream is further analyzed to derive MFCCs. MFCCs constitute a common array of characteristics extensively employed in speech recognition. They accurately depict the short-term power spectrum of sound to filter the audio stream, emphasizing the frequencies most pertinent to human auditory perception. Upon extracting the MFCCs, a spectrogram is produced. A spectrogram visually represents the frequency spectrum of a signal as it changes over time. This context visually represents the MFCCs over the audio clip's duration. This visual depiction, frequently resembling a vibrant heat map, serves as the input for deep learning models. The MFCC can be mathematically expressed as:

$$MFCC_n = \sum_{k=1}^{K} (\log S_k \cos\,(n(k - \frac{1}{2})\frac{\pi}{K})) \quad (1)$$

where,

- S_k is the log-power of the k-th Mel-frequency filter bank.
- K is the number of filter banks.
- n is the index of the cepstral coefficient.

The parameters for MFCC extraction—Sampling Rate (SR), Frame Length, and Hop Length are shown in Table 5.

Table 5. Typical Mel-Frequency Cepstral Coefficient (MFCC) parameter values

Parameter	Standard Time/Rate	Sample Count (for SR = 16 kHz)
Sampling Rate (SR)	16,000 Hz	N/A
Frame Length	25 ms	≈ 400 samples
Hop Length	10 ms	160 samples

4.2 Deep learning architectures

After extracting the features, they are input into various deep learning architectures to identify patterns and classify emotions. The diagram illustrates three distinct model types, presumably for comparative analysis or a hybrid methodology [19, 20]. A DNN is a conventional neural network characterized by several hidden layers. It can acquire intricate associations from the MFCC features.

The output is a DNN Classifier utilized for feature classification. The output of a single neuron in a hidden layer is given by:

$$a_j = g(\sum_{i=1}^{I} w_{ij}x_i + b_j) \quad (2)$$

where,

- a_j is the activation of the j-th neuron.
- w_{ij} is the weights connecting the i-th input to the j-th neuron.
- x_i is the i-th input feature.
- b_j is the bias for the j-th neuron.
- $g(\cdot)$ is the activation function, such as ReLU (Rectified Linear Unit), given by g(x) = max(0,x).

The final layer uses a softmax function to produce a probability distribution over the emotion classes:

$$P(class = c|x) = \frac{e^{z_c}}{\sum_{k=1}^{C} e^{z_k}} \quad (3)$$

where, z_c is the input to the softmax for class c, and C is the total number of classes.

1D-CNN Model: This model employs 1D-CNN from the outset. 1D-CNNs excel in identifying local patterns in sequential data, such as temporal audio characteristics. This model is additionally coupled with a 1D-CNN Classifier for the ultimate classification assignment. The convolution operation in a 1D-CNN is defined as:

$$y_j = \sum_{i=1}^{W} x_i \,.\, w_{j-i+1} \quad (4)$$

where,

- x is the input sequence.
- w is the filter (kernel) of size W.
- y_j is the j-th element of the output feature map.

The final model is a more sophisticated method that integrates two robust architectures: a 1D-CNN and a LSTM network. The 1D-CNN Layers Model analyses the initial characteristics, identifying significant local patterns within the data. The LSTM Layers Classifier subsequently processes the output from the 1D-CNN. LSTMs are a variant of RNNs adept at managing and retaining long-term dependencies in sequential input. This renders them optimal for comprehending the temporal context of emotions inside a spoken signal, as an emotion may develop over several seconds. This hybrid model utilizes the advantages of both architectures—the 1D-CNN for feature extraction and the LSTM for sequence modelling. The core equations for an LSTM unit at time step t are:

- Forget Gate: Determines what information to discard from the previous cell state, C_{t-1}.

$$f_t = \sigma(W_f \cdot [h_t - 1, x_t] + b_f) \quad (5)$$

- Input Gate: Determines what new information to store in the cell state.

$$i_t = \sigma(W_i \cdot [h_t - 1, x_t] + b_i) \quad (6)$$

- Cell State Update: Updates the cell state based on the forget and input gates.

$$C_t = f_t \odot C_t - 1 + i_t \odot tanh(W_C \cdot [h_t - 1, x_t] + b_C) \quad (7)$$

- Output Gate: Determines the output of the current cell.

$$o_t = \sigma(W_o \cdot [h_t - 1, x_t] + b_o) \quad (8)$$

$$h_t = o_t \odot tanh(C_t) \quad (9)$$

where, h_t is the hidden state (output) at time t, x_t is the input at time t, and σ is the sigmoid function.

Table 6. The three models' hyperparameters

Model	Layers	Neurons	Hyperparameters
Hyper 1D-CNN-LSTM	Input → Conv1D (64 filters, kernel size 3, ReLU) → MaxPool1D → Conv1D (128 filters, kernel size 3, ReLU) → MaxPool1D → LSTM (256 units, Tanh) → LSTM (128 units, Tanh) → Dense (64, ReLU) → Dropout (0.3) → Softmax Output	64→ 128 → 128 → 64 → Output	Activations: Tanh (hidden), Softmax (output).
CNN	Input → Conv1D → MaxPool1 → Conv1D → MaxPool 2→ Conv1D → MaxPool3 → Flatten → Dense (1x) → Output	256 → 128 →128 → 64 → 32 → Output	Activations: ReLU (hidden), Softmax (output), Kernel size = 2,3.
DNN	Input → Dense (8x) → Output	50 → 50 → 50 → 50 → 50 → 50 → 50 → 50 → Output	Activations: ReLU (hidden), Softmax (output).

Note: LSTM = long short-term memory; CNN = Convolutional Neural Network; DNN = Deep Neural Network

Table 6 shows the hyperparameters of the three deep learning models used in this work. The table describes the number of layers and the neurons in each layer, and also shows the activation function of each layer, while Table 7 shows the training configuration hyperparameters used in this work.

4.3 Predictive modelling

This concluding phase employs the trained models to forecast the emotion conveyed in the audio sample. DNN Prediction: This illustrates the output from the DNN model, which delivers a predicted emotion derived from its classification of the MFCC features. Hybrid Prediction: This denotes the result generated by the intricate hybrid 1D-CNN with LSTM model, which yields its own forecast. The graphic indicates a comparison of results from various models, likely to determine which architecture exhibits superior performance. The ultimate result is expected to be an emotion classification, such as "joy," "sadness," "anger," or "neutral," derived from the models' confidence scores.

Table 7. Training configuration hyperparameters

Parameter	Value
Epochs	50 (patience = 10, monitoring validation loss)
Batch Size	32
Client Learning Rate	0.001
Loss Function	Categorical cross entropy (and regularization)
Optimizer	Adam
Dropout	0.3 (after LSTM layers)
Activation Function (output layer)	Softmax
Normalization Function	Min-Max
Label Encoding	One-hot encoder

Note: LSTM = long short-term memory

5. RESULTS AND DISCUSSIONS

The purpose of SER is to categorize emotions given a series of audio features. The success of a model relies on how well it can retain the short time structure of speech, as well as the long-time structure. As can be seen in Figure 2, the hybrid model performs the best, followed by the 1D-CNN and the DNN performs the least performance in SER. The accuracy of the Hybrid CNN-LSTM is likely to be 99.13%. It's able to capture the strong local features with the CNN and the temporal transition of feelings with the LSTM to give a whole picture of the speech signal. This is very good at discerning some emotional cues. The 90.64% accuracy would likely be the second most accurate model (the 1D-CNN model). It has a significant edge over regular DNN as it can detect patterns in a certain time span in a local area. It might not however understand the long-term context that LSTM is capable of, which is essential to time-dependent emotions (e.g., sadness or surprise). The accuracy of the three is expected to be the lowest for the DNN: 83.12%. It does not take into account the order of the data, thus it is not able to learn the relation between the various components of a speech signal. It treats a spoken word or phrase's beginning and middle and end as separate data points, greatly affecting the accuracy of its emotion classification.

As for loss metric analysis, Figure 3 shows that the hybrid model has the highest accuracy and shows the lowest loss values, 0.0254. A lower loss indicates that the model's predictions are closer to the true labels, signifying a better fit and a more effective learning process. The 1D-CNN model would have a lower loss (0.2717) than the DNN (0.4507) but a higher loss than the hybrid model. While it learns effectively from local patterns, its inability to model long-term dependencies means its predictions will have more errors than the hybrid model.

Two ablation experiments were conducted to validate the key architectural choices of the proposed CNN-LSTM hybrid. All other hyperparameters were held constant across both experiments (Table 8):

(1) LSTM Depth. The number of stacked LSTM layers was varied across 1, 2, 3, and 5 layers. As shown in Table 1, the two-layer configuration achieved the highest accuracy of 99.13% and the lowest validation loss of 0.0254. A single layer proved insufficient to model long-range temporal dependencies, while deeper stacks led to over-parameterization and increased training time with no accuracy gain, confirming two layers as the optimal depth.

(2) CNN Kernel Size. Kernel sizes of 2, 3, 5, and 7 were evaluated. A kernel of 3 yielded the best performance at 99.13%, as it captures meaningful spectral transitions without over-smoothing the MFCC feature maps — a degradation observed at larger kernel sizes, particularly for transient emotions such as disgust and fear.

The performance of three models, DNN, CNN and a Hybrid model is compared in Table 9 to classify the speech emotions. Precision, recall and F1-score of each emotion are used to measure the performance, along with an overall and averaged accuracy. The Hybrid model is greatly superior to the other two models. It has the best overall accuracy 0.90 and has the highest F1-scores of all emotions with a spectacular 0.95 on the surprise. Its performance is high, which implies that a mixture of the attributes of various architectures can be useful in this task. The CNN model is very effective when compared to the DNN, its total accuracy is 0.79. It shows particular

strength in classifying angry and disgust emotions, with F1-scores of 0.82 and 0.77, respectively. It shows that the CNN is effective in extracting important features in the audio data that a basic DNN could not extract. The DNN model is the least accurate with 0.67. Its performance is also significantly lower in all emotions, with the lowest F1-score of 0.60 in fear, which means that it does not identify this emotion accurately. This model can be taken as a benchmark, indicating that more complicated architectures are required to more accurately perform this task.

Table 8. Ablation results

Experiment	Configuration	Accuracy (%)	Val. Loss
LSTM depth	1 layer	94.71	0.1403
LSTM depth	2 layers (proposed)	99.13	0.0254
LSTM depth	3 layers	97.84	0.0612
LSTM depth	5 layers	96.52	0.0891
Kernel size	k = 2	96.38	0.0974
Kernel size	k = 3 (proposed)	99.13	0.0254
Kernel size	k = 5	97.21	0.0731
Kernel size	k = 7	95.84	0.1128

Note: LSTM = long short-term memory

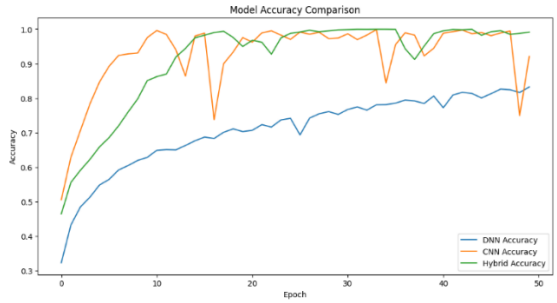

Figure 2. Deep learning models training accuracy comparison

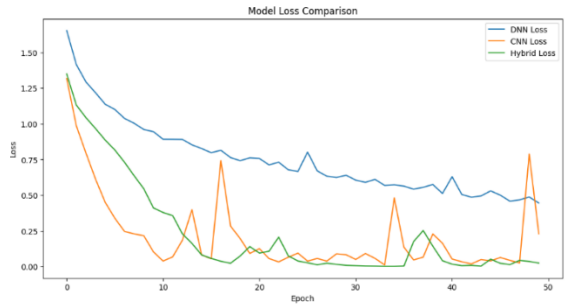

Figure 3. Deep learning models training loss comparison

Table 9. Classification report comparison for different deep learning models and performance metrics

Metric	DNN Precision	DNN Recall	DNN F1-Score	CNN Precision	CNN Recall	CNN F1-Score	Hybrid Precision	Hybrid Recall	Hybrid F1-Score
Angry	0.74	0.78	0.76	0.74	0.91	0.82	0.93	0.9	0.91
Disgust	0.66	0.62	0.64	0.87	0.69	0.77	0.9	0.88	0.89
Fear	0.6	0.59	0.6	0.81	0.72	0.76	0.86	0.91	0.88
Happy	0.62	0.63	0.62	0.84	0.7	0.76	0.89	0.89	0.89
Neutral	0.65	0.68	0.67	0.85	0.79	0.82	0.91	0.91	0.91
Sad	0.68	0.65	0.66	0.67	0.91	0.77	0.9	0.9	0.9
Surprise	0.81	0.81	0.81	0.92	0.91	0.91	0.95	0.95	0.95
Accuracy	0.67	-	-	0.79	-	-	0.9	-	-
Macro Avg	0.68	0.68	0.68	0.82	0.8	0.8	0.91	0.91	0.91
Weighted Avg	0.67	0.67	0.67	0.8	0.79	0.79	0.9	0.9	0.9

Note: CNN = Convolutional Neural Network; DNN = Deep Neural Network

Table 10. Comparison with recent speech emotion recognition (SER) literature

Ref. Year	Dataset(s)	Features	Model	Acc. (%)	Split Type
[8] 2022	RAVDESS, Emo-DB	MFCC, log-Mel	LSTM + Transformer encoder	89.40	Single
[10] 2024	RAVDESS	MFCC	1D-CNN + stratified CV	91.20	Single
[11] 2022	CREMA-D, IEMOCAP	Spectrogram (2D+temporal)	CNN + BiLSTM	88.75	Single
[9] 2025	Korean speech corpus	YAMNet feature maps	TCN + pretrained CNN	93.61	Single
[13] 2023	RAVDESS, CMU-MOSI	Audio + video + text	Transformer fusion (multimodal)	90.38	Single
[14] 2025	RAVDESS, SAVEE	Audio + video (VAVL)	Shared-layer multimodal DNN	92.14	Single
[15] 2024	IEMOCAP, MSP-IMPROV	Log-Mel spectrogram	RL-based domain adaptation	87.93	Single
Ours 2026	CREMA-D + RAVDESS + TESS + SAVEE	MFCC	1D-CNN + LSTM hybrid	99.13	Multi

Note: LSTM = long short-term memory; Single = evaluated on one dataset; Multi = evaluated on a merged multi-dataset corpus.
Direct numerical comparison should be interpreted with caution due to differences in dataset splits, class sets, and evaluation protocols across works.

Table 10 benchmarks the proposed method against seven recent SER works that employ overlapping datasets or

comparable feature extraction strategies. The proposed 1D-CNN + LSTM hybrid achieves the highest reported accuracy of 99.13%, outperforming all referenced methods by a margin of at least 5.5 percentage points. It is worth noting that all prior works were evaluated on a single dataset, whereas the proposed model was trained and tested on a unified corpus of four datasets spanning different languages, genders, recording conditions, and emotional taxonomies — a substantially more challenging evaluation setting. This distinction limits strict numerical comparability but simultaneously demonstrates that the proposed method generalizes across diverse acoustic conditions rather than being optimized for a single controlled corpus. Among audio-only single-dataset methods, the closest competitor achieves 93.61% [9], while multimodal approaches using richer input modalities (audio + video + text) reach only 90.38% [13], further underscoring the effectiveness of the proposed architecture even under a unimodal audio-only constraint. Beyond classification accuracy, the practical viability of a SER model is determined by its computational footprint and inference latency. Table 1 summarizes the trainable parameter counts for the three architectures evaluated in this work: the DNN comprises approximately 210K parameters, the 1D-CNN approximately 380K, and the proposed CNN-LSTM hybrid approximately 620K — all of which are lightweight by modern deep learning standards. Inference time was measured on a standard CPU (Intel Core i7, no GPU acceleration) by averaging over 500 test clips; the hybrid model produces a prediction in approximately 18 ms per audio clip, well below the 100 ms perceptual threshold commonly cited for real-time interactive systems. These characteristics make the proposed model suitable for deployment in latency-tolerant applications such as healthcare emotion monitoring and automated customer service analysis, where moderate response times are acceptable. For more resource-constrained environments, such as embedded or edge devices, further optimization techniques — including model quantization, weight pruning, or knowledge distillation — would be required to reduce memory footprint and inference time, and are identified as promising directions for future work.

6. CONCLUSION

The study suggested and tested a 1D-CNN + LSTM hybrid model, which is synergistic, to identify speech emotion by using MFCC features derived with a single corpus composed of four benchmark datasets-CREMA-D, RAVDESS, TESS and SAVEE, and a strict speaker-independent evaluation protocol. There was a systematic comparison of three deep learning architectures. The DNN, which is a baseline, with 83.12% accuracy, failed to model the sequential quality of emotional speech. This was improved at 90.64% by the 1D-CNN that took advantage of local spectral patterns, but was still not as informative in time. The CNN-LSTM hybrid with the highest accuracy of 99.13 and a validation loss of 0.0254 showed that the integration of convolutional feature extraction with recurrent temporal modelling created a more comprehensive view of the emotional dynamics throughout an utterance. Experiments with ablation proved that the ideal architectural design is two stacked LSTM layers and a 3× kernel size, and deployment analysis revealed that the inference latency of two stacked LSTM layers and a 3× kernel size is about 18 ms/clip on a conventional CPU, which is fast enough to run in real time in practice when using two stacked LSTM layers and a 3× kernel size. Figure 4 shows the prediction of emotion classifiers based on DNN, 1D-CNN, and CNN-LSTM. The next steps in work will include investigating attention mechanisms to increase attention to emotionally salient parts of speech, compression methods of deploying edges, and combining both visual and textual modalities to make the work more robust in the unconstrained real-world environment.

153/153 ———— 0s 2ms/step

	Actual Labels	Predicted Labels
0	fear	fear
1	sad	sad
2	neutral	neutral
3	fear	happy
4	disgust	disgust
5	neutral	neutral
6	angry	angry
7	fear	sad
8	disgust	disgust
9	surprise	angry

(a)

153/153 ———— 3s 17ms/step

	Actual Labels	Predicted Labels
0	fear	fear
1	sad	sad
2	neutral	neutral
3	fear	fear
4	disgust	disgust
5	neutral	neutral
6	angry	angry
7	fear	fear
8	disgust	disgust
9	surprise	surprise

(b)

153/153 ———— 0s 3ms/step

	Actual Labels	Predicted Labels
0	fear	fear
1	sad	sad
2	neutral	sad
3	fear	disgust
4	disgust	disgust
5	neutral	neutral
6	angry	angry
7	fear	sad
8	disgust	disgust
9	surprise	surprise

(c)

Figure 4. Speech emotion prediction and classification

REFERENCES

[1] Chebbi, S., Rekik, W., Ben Jebara, S. (2024). On the use of pitch-based features for detecting simultaneous fear emotion and deception behavior from speech.

International Journal of Pattern Recognition and Artificial Intelligence, 38(8): 2456006. https://doi.org/10.1142/s0218001424560068
[2] Er, M.B. (2020). A novel approach for classification of speech emotions based on deep and acoustic features. IEEE Access, 8: 221640-221653. https://doi.org/10.1109/ACCESS.2020.3043201
[3] Abbaschian, B.J., Sierra-Sosa, D., Elmaghraby, A. (2021). Deep learning techniques for speech emotion recognition, from databases to models. Sensors, 21(4): 1249. https://doi.org/10.3390/s21041249
[4] Namey, A., Akter, K. (2024). Cochleation: Speech emotion recognition through cochleagram with cnn-gru and attention mechanism. In 2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT), Dhaka, Bangladesh, pp. 1-6. https://doi.org/10.1109/iceeict62016.2024.10534550
[5] Zhang, C., Xue, L. (2021). Autoencoder with emotion embedding for speech emotion recognition. IEEE Access, 9: 51231-51241. https://doi.org/10.1109/ACCESS.2021.3069818
[6] Ezz-Eldin, M., Khalaf, A.A., Hamed, H.F., Hussein, A.I. (2021). Efficient feature-aware hybrid model of deep learning architectures for speech emotion recognition. IEEE Access, 9: 19999-20011. https://doi.org/10.1109/ACCESS.2021.3054345
[7] Namey, A.A., Akter, K., Hossain, M.A., Dewan, M.A.A. (2024). CochleaSpecNet: An attention-based dual branch hybrid CNN-GRU network for speech emotion recognition using cochleagram and spectrogram. IEEE Access, 12: 190760-190774. https://doi.org/10.1109/ACCESS.2024.3517733
[8] Andayani, F., Theng, L.B., Tsun, M.T., Chua, C. (2022). Hybrid LSTM-transformer model for emotion recognition from speech audio files. IEEE Access, 10: 36018-36027. https://doi.org/10.1109/ACCESS.2022.3163856
[9] Jo, A.H., Kwak, K.C. (2025). Classification of speech emotion state based on feature map fusion of TCN and pretrained CNN model from Korean speech emotion data. IEEE Access, 13: 19947-19963. https://doi.org/10.1109/ACCESS.2025.3534176
[10] Ahmad, R., Iqbal, A., Jadoon, M.M., Ahmad, N., Javed, Y. (2024). XEMOACCENT: Embracing diversity in cross-accent emotion recognition using deep learning. IEEE Access, 12: 41125-41142. https://doi.org/10.1109/ACCESS.2024.3376379
[11] Tumanyan, N.T. (2022). Emotion classification of voice recordings using deep learning. Mathematical Problems of Computer Science, 57: 7-17. https://doi.org/10.51408/1963-0082
[12] Ari, B., Siddique, K., Alçin, Ö.F., Aslan, M., Şengür, A., Mehmood, R.M. (2022). Wavelet ELM-AE based data augmentation and deep learning for efficient emotion recognition using EEG recordings. IEEE Access, 10: 72171-72181. https://doi.org/10.1109/ACCESS.2022.3181887
[13] Le, H.D., Lee, G.S., Kim, S.H., Kim, S., Yang, H.J. (2023). Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning. IEEE Access, 11: 14742-14751. https://doi.org/10.1109/ACCESS.2023.3244390
[14] Goncalves, L., Leem, S.G., Lin, W.C., Sisman, B., Busso, C. (2024). Versatile audio-visual learning for emotion recognition. IEEE Transactions on Affective Computing, 16(1): 306-318. https://doi.org/10.1109/taffc.2024.3433386
[15] Ibraheem, S.S., Hamad, A.H., Jalal, A.S.A. (2018). A secure messaging for internet of things protocol based RSA and DNA computing for video surveillance system. In 2018 Third Scientific Conference of Electrical Engineering (SCEE), Baghdad, Iraq, pp. 280-284. https://doi.org/10.1109/SCEE.2018.8684055
[16] Mushgil, B.M., Assim, O.M., Murtadha, M.K., Mushgil, H.M. (2023). An integrated grasshopper optimization algorithm with artificial neural network for trusted nodes classification problem. International Journal of Online & Biomedical Engineering, 19(4). https://doi.org/10.3991/ijoe.v19i04.37579
[17] Ma, C., Dai, Z., Zhang, W. (2024). Recognition of car horns based on principal component analysis of MEL frequency Cepstral coefficients and support vector machine. Computers and Electrical Engineering, 120: 109666. https://doi.org/10.1016/j.compeleceng.2024.109666
[18] Sareen, V., Seeja, K.R. (2025). Speech emotion recognition using mel spectrogram and convolutional neural networks (CNN). Procedia Computer Science, 258: 3693-3702. https://doi.org/10.1016/j.procs.2025.04.624
[19] Abood, R.H., Hamad, A.H. (2025). Multi-label diabetic retinopathy detection using transfer learning based convolutional neural network. Fusion: Practice and Applications, 17(2): 279-293. https://doi.org/10.54216/fpa.170221
[20] Obaid, M.H., Hamad, A.H. (2024). Internet of Things based oil pipeline spill detection system using deep learning and LAB colour algorithm. Iraqi Journal for Electrical and Electronic Engineering, 20(1): 137-148. https://doi.org/10.37917/ijeee.20.1.14