**Bulletin of Electrical Engineering and Informatics**
Vol. 14, No. 1, February 2025, pp. 316~327
ISSN: 2302-9285, DOI: 10.11591/eei.v14i1.8464 

# Analysis of human emotions through speech using deep learning fusion technique for Industry 5.0

**Chevella Anil Kumar, Vumanthala Sagar Reddy, Ambati Pravallika, Rao Y. Chalapathi, Neelam Syamala**
Department of Electronics and Communication Engineering, Vallurupally Nageshwara Rao Vignana Jyothi Institute of Engineering and Technology, Hyderabad, Telangana, India

**Article Info**

***Article history:***

Received Mar 14, 2024
Revised Oct 5, 2024
Accepted Oct 17, 2024

***Keywords:***

Artificial intelligence
Convolutional neural network
Emotion recognition
Long short-term memory
Ryerson audio-visual database of emotional speech and song

**ABSTRACT**

Emotions are important for human well-being and social connections. This work focuses on the issue of effectively understanding emotions in human speech, specifically in the context of Industry 5.0. Traditional approaches and machine learning (ML) techniques for identifying emotions in speech are limited, such as the requirement for complicated feature extraction. Traditional methods yield recognition accuracies of no more than 90% because to the restricted extraction of temporal/sequence information. This paper suggests a ground-breaking fusion-based deep learning (DL) method to overcome these limitations. Specifically, one-dimensional (1D) and two-dimensional (2D) convolution neural network (CNN) can automatically extract significant characteristics and handle enormous datasets in real time. Furthermore, a fusion-based DL network, speech emotion recognition deep learning fusion network (SER_DLFNet), has been proposed, which combines CNN with long short-term memory (LSTM) to collect sequence information and increase recognition accuracy. The proposed model shows impressive results, with a test accuracy of 95.52% on the ryerson audio-visual database of emotional speech and song (RAVDESS) dataset. This research contributes to the advancement of more precise and efficient emotion identification algorithms for voice analysis, especially within the framework of Industry 5.0.



***Corresponding Author:***

Vumanthala Sagar Reddy
Department of Electronics and Communication Engineering
Vallurupally Nageshwara Rao Vignana Jyothi Institute of Engineering and Technology
Hyderabad, Telangana, India
Email: vsagarreddy1990@gmail.com

## 1. INTRODUCTION

The goal of speech emotion recognition (SER) is to develop automatic systems that can recognise and interpret speech signals that represent different emotional states, including fear, surprise, disgust, rage, happiness, sadness, and neutrality. Numerous applications exist for SER, including mental health assessment and human-computer interaction. Additionally, they can design more efficient virtual assistants that can modify their responses to reflect the user's attitude. The emotional content of speech can also be used to identify mental diseases. It can also be applied in market research to find out how people feel about advertisements or messaging about products. In the SER context, deep learning (DL) has become more important, as it can learn features from raw speech data automatically, which helps to improve emotion recognition accuracy.

Research has demonstrated that DL models exhibit greater resilience to noise and voice signal variance than typical machine learning (ML) techniques. Additionally, they perform better because they scale well with huge datasets, which improves generalisation for DL feature extraction, numerous network models and architectures are feasible. Figure 1 illustrates the proposed SER system, which features our developed model, speech emotion recognition deep learning fusion network (SER_DLFNet) - a novel, customized DL network that utilizes fusion-based architecture.

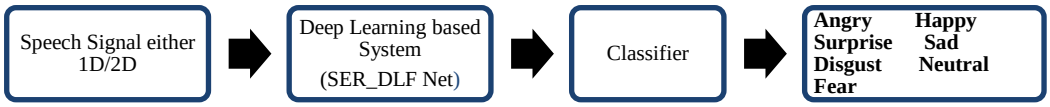


Figure 1. SER system

To identify the emotions expressed in speech data, we used a fusion based DL network model SER_DLFNet in this suggested framework, which combines one- and two-dimensional (1D and 2D) convolution neural network (CNN)+long short-term memory (LSTM) architectures. When handling 1D time-series data, such audio signals in the speech emotion detection task, 1D convolutional neural networks (1D-CNNs) are very helpful. 1D-CNNs assist in finding patterns and traits that are relevant for emotion recognition by effectively extracting relevant features from audio signals [1]. In contrast, 2D convolutional neural networks (2D-CNNs) can analyse 2D input data, including images, and are especially useful for analysing spectrograms, also known as Mel-spectrograms, which are frequently used in SER applications. Spectrograms can be viewed as visuals and are used to represent audio signals in the time-frequency domain. Thus, 2D-CNNs demonstrate that they are effective in examining spectrograms for emotion recognition, as they have demonstrated a propensity for producing accurate outcomes [2], [3]. In SER tasks, 1D-CNNs and 2D-CNNs [3] are both important. The former is used for extracting features from raw audio signals, while the latter is used for spectrogram or Mel-spectrogram analysis. The overall efficacy of the framework is enhanced by the fusing of these CNNs with the LSTM, as suggested in fusion-based network model SER_DLFNet.

## 2. LITERATURE

Speech is among the most popular and organic forms of interpersonal communication. It conveys a lot about the speaker's thoughts, feelings, and goals. To develop a speech emotion detection system, it is essential to surmount the challenging challenge of identifying and extracting emotion-related data from speech [4]. SER is based on linear prediction cepstrum coefficients, fundamental frequency (F0), and Mel frequency cepstral coefficiens (MFCCs), which are commonly employed for emotion recognition from speech. Various methodologies were utilized to extract emotion from speech, with each study employing a unique array of speech variables [5]. In recent years, DL algorithms [6] have proven to be quite helpful in this field. Affective computing holds that for machines to function well, they need to be able to identify emotions. For instance, using robots to assist with elderly care or as hospital porters requires a high level of environmental awareness. Human expressions on the face and in voice convey a person's inner feelings [7]–[9]. DL has the potential to greatly improve natural language communication as well as human-machine interactions.

The extraction and classification of features are the two most important phases in the SER process. Many properties, including as prosodic traits, source-based excitation features, and linguistic elements, were found at the initial stages of speech processing. For feature classification, both linear and nonlinear classifiers are used in the second step. The maximum-likelihood principle, Bayes networks, and support vector machine (SVM) models are the three most commonly utilized classification techniques for emotion recognition [10]. It's common to think of speech as a non-stationary signal. Non-linear classifiers [11] are best option for classification. Gaussian mixture model (GMM) and the hidden Markov model are examples of non-linear classifiers [11]–[13]. Classical ML techniques based on GMM models, such as SVMs [12], [14], [15] artificial neural networks (ANNs), and GMM models have shown excellent success in SER tasks. However, there are inherent limits to the ability of technologies to accurately identify emotional states in speech. ML and artificial intelligence (AI) algorithms are putting a lot of effort into improving the accuracy of voice-based emotion recognition. Unlike traditional methods, DL algorithms extract features without human intervention, in contrast to conventional approaches.

DL algorithms have been proposed as a solution to the accuracy issues with regular ML algorithms. The intelligent DL environment analyses human emotions using voice analysis using a CNN [17], and LSTM [16], [17]. Numerous research projects on DL for speech-based emotional analysis have been carried out in

light of the AI technology's explosive rise. At the moment, some progress has been made in identifying emotions using DL techniques based on speech. For instance, utilizing DL and cognitive wireless frameworks, Hossain and Muhammad [18] created an audio-visual emotion identification system [19]. This technology was able to automatically interpret a patient's emotional state by looking at their facial expressions. After putting the strategy to the test, they were able to show that it would help the expansion of online healthcare. Khalil *et al.* [4] offered additional details and a brief synopsis of the research on DL for SER in 2019. Ntalampiras [20] used behavioral analogies to construct a twin neural network for voice emotion recognition.

A unique temporal modelling framework for robust emotion categorization was presented by Zheng *et al*. [21]. It incorporates CNN, capsule networks, and a bidirectional LSTM network (BLSTM) [22], [23]. The goal of this approach is to effectively manage speech signals' temporal dynamics and to give a cutting-edge technique for identifying the extracted patterns, with an accuracy of 69.40%. A novel framework for SER was presented by Mustaqeem *et al*. [22]. It selects crucial sequence segments by evaluating the similarity measures of radial basis function networks (RBFNs) in clusters. The short-time Fourier transform (STFT) algorithm is then used to convert the chosen sequence into a spectrogram, which is subsequently input into a CNN to extract significant and discriminative characteristics from the spoken spectrogram. After the features are standardised, a deep BiLSTM network is trained with them. This network learns the temporal information needed to identify the final emotion state with 91.14% accuracy.

## 3. METHOD

The entire architecture of the proposed network, which consists of LSTM, 1-D and 2-D CNN, is depicted in Figure 2. We proposed a fusion based DL network model (SER_DLFNet) that includes a local feature learning block (LFLB) shown in Figure 3 comprising a convolutional layer, Max-pooling, exponential linear unit (ELU) and batch normalization (BN) layer for extracting local features. Additionally, the model integrates an LSTM network to capture long-term dependencies from the sequence of local features.

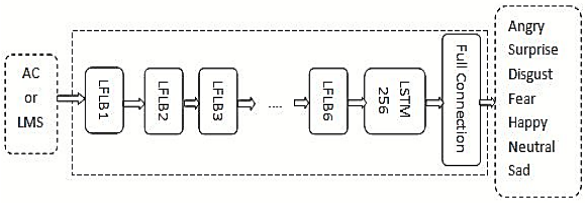


Figure 2. Architecture of SER_DLFNet, **audio clips (1D signal)=AC, *log Mel-spectrogram (2D signal)=LMS*

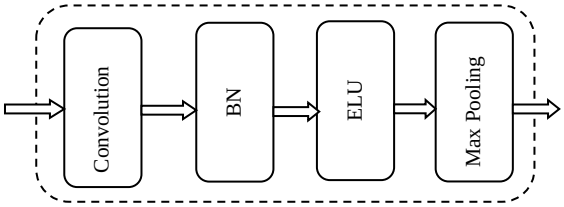


Figure 3. LFLB

### 3.1. Deep learning of features

Combining LFLB and LSTM allows for the learning of both local and global characteristics from raw audio samples. The core component of LFLB is the convolution layer, which is made to analyse data in a grid format. It is capable of learning a sequence feature in which every feature member is determined by a limited number of nearby input members. Every feature of the learnt feature is a function of the output's prior features, in contrast to LSTM, which is designed to handle a series of values. Thus, by combining the LSTM and CNN, we may learn high-level features that incorporate both local and long-term contextual dependencies.

#### 3.1.1. Acquisition of local features

The LFLB shown in Figure 3 incorporates essential layers like convolution and pooling, with convolution emphasizing connectivity and weight sharing. The BN layer enhances deep neural network performance and stability by normalizing activations in each batch. This helps maintain mean and standard deviation close to 0 and 1 respectively. The BN layer output is specified by the ELU layer, which, unlike most activation functions, permits negative values, which approaches 0 for mean activations for faster learning and improved accuracy. The pooling layer strengthens features, making them less susceptible to distortion and noise; the most widely utilized nonlinear function is max-pooling that outputs the greatest values from sub-regions. For 1D input signal, the output of CNN in LFLB calculated as (1):

$$q(n) = p(n) * w(n) = \sum_{m=-l}^{l} p(m).w(n-m) \quad (1)$$

where $q(n)$ represents the output of the CNN layer for an input signal $p(n)$, obtained by convolving using a Kernel $w(n)$ of size $n$.

On the other hand, in the case of 2D-CNN, where $p(i,j)$ input to the 2D convolution layer, the output $q(i,j)$ can be obtained by convolving $p(i,j)$ with the Kernel $w(i,j)$.

$$q(i,j) = p(i,j) * w(i,j) = \sum_{s=-l}^{l} \sum_{t=-m}^{m} p(s,t).w(i-s, j-t) \quad (2)$$

Next, it normalizes the activations of the previous layer at each batch by feeding the convolved features into the BN layer. The convolved characteristics variance and mean are maintained by the BN layer near one and zero, respectively. Following the normalized features application of the ELU layer, the final features are stated as (3):

$$f_i^l = \sigma(BN(b_i^n + \sum_j f_j^{\mathrm{n}-1} * w_{ij}^n) \quad (3)$$

where $f_i^l$ and $f_j^{\mathrm{n}-1}$denote the $n^{th}$ output feature at the $n^{th}$ layer and the $j^{th}$ input feature at the $(\mathrm{n}-1)^{th}$ layer, respectively; $w_{ij}^n$ represents the convolution Kernel between the $i^{th}$ and $j^{th}$ features. The normalization features achieved using the function BN $(\cdot)$. The ELU activation function can be expressed as (4):

$$\sigma(x) = \begin{cases} x^{x \geq 0} \\ \alpha(e^x - 1)_{x<0} \end{cases} \quad (4)$$

Euler's number, $e$, implies there needs to be a positive extra alpha constant (> 0). The pooling layer performs the down-sampling function, which reduces the resolution of the feature. For a given pooling region with index $k$, the input feature of the lth max-pooling layer, denoted as $h_p^l$, transforms into the output feature of the lth pooling layer $h_k^l$. Features of the max-pooling layer are as (5):

$$h_k^l = \underset{\forall p \in \Omega_k}{max} h_p^l \quad (5)$$

### 3.1.2. Global feature learning

The LSTM architecture is employed to capture prolonged dependencies among sequences, and it is layered on top of the LFLB to grasp contextual dependencies from the acquired local feature sequences. The LSTM efficiently incorporates or discards information from the block state by utilizing input, output, forget, and cell states. Multiple equations describe the procedure for updating an LSTM unit [24], [25]. Expressing the connection between the inputs and outputs of an LSTM unit can be done as (6) to (11):

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i) \quad (6)$$

$$\hat{c}_t = tanh(w_c[h_{t-1}, x_t] + b_c) \quad (7)$$

$$c_t = f_t * c_{t-1} + i_t * \hat{c}_t \quad (8)$$

$$f_t = \sigma(w_f.[h_{t-1}, x_t] + b_f) \quad (9)$$

$$o_t = \sigma(w_o.[h_{t-1}, x_t] + b_o) \quad (10)$$

$$h_t = o_t * tanh(c_t) \quad (11)$$

In the context where ct signifies the LSTM unit state, w and b are parameter matrices and vectors, $f_t$, it, and $o_t$ represent gate vectors, $\sigma$ denotes the sigmoid function, and $C$ and $h$ are hyperbolic tangents, the Hadamard product is denoted by $*$.

### 3.2. Dataset

A public repository of recordings of emotional speech and song is RAVDESS. A range of emotions, including happy, neutral, sad, surprised, fearful, angry, and disgusted, are portrayed by 24 professional performers, both in speech and song. The actors, who are fluent in English, were asked to record the different emotions. For every emotion, three types of modalities are offered: audio-only, video-only, and audio-video. To train our model to identify and categorise various emotions in speech data, we utilised a dataset that contained only audio.

The audio-only speech dataset consists of 1440 ".wav" files, each having a resolution of 16 bits and a sampling rate of 48 kHz. These files constitute a total of 1440 trials, distributed among 24 performers, with each performer contributing 60 trials. The dataset includes 60 speech recordings for each emotion, featuring

30 recordings by female speakers and 30 by male speakers. Each recording involves two statements: 01=dogs are sitting by the door and 02=kids are talking by the door [26]. Here, the Figures 4 and 5 display various emotional speech files along with their corresponding spectrograms from the RAVDESS dataset.

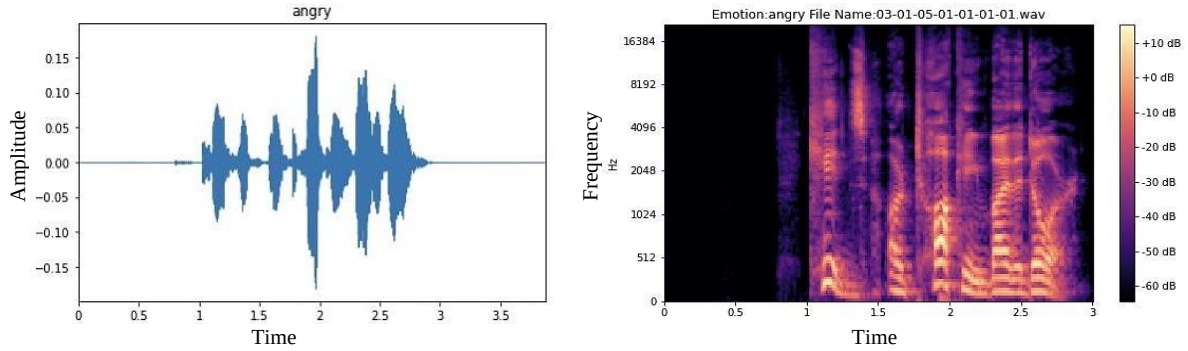


Figure 4. Angry emotional speech signal along with their corresponding spectrograms

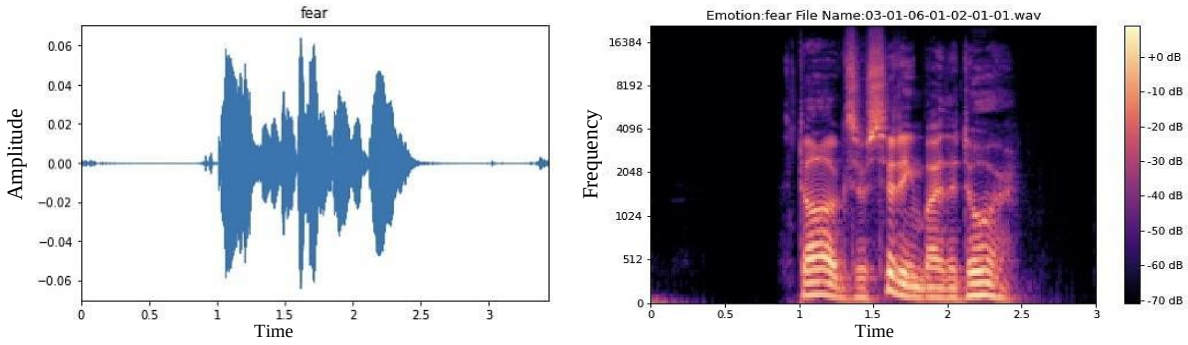


Figure 5. Fear emotional speech signal along with their corresponding spectrogram

## 4. EXPERIMENTAL STUDIES

Before building the proposed fusion based network model using 1D-CNN and 2D-CNN, we trained and evaluated the emotional voice dataset, comparing their accuracies to the proposed network model SER_DLFNet which consists of 1D and 2D CNNs+LSTM. The 1D network processes 1D audio vectors, whereas the 2D network processes spectrograms. Figures 6 and 7 shows how the LFLB collects local features, which are then input into the LSTM layer to capture global features and contextual dependencies. The LSTM output includes both immediate and a long-term background data. An additional fully interconnected layer connects directly to the LSTM layer.

$$z^l = b^t + z^{l-1}.w^t \tag{12}$$

Lastly, a softmax layer is employed to classify the emotion of the input data, determining the prediction probability for each class. The class label y has various potential interpretations. The Softmax function can be defined as in (14), and ultimately, the predicted class label $\tilde{y}$ is computed using (15).

$$z_i = \sum w_{ji} * h_j \tag{13}$$

$$Softmax(z_i) = p_i = \frac{e^{z_i}}{\sum_{j=0}^{k} e^{z_j}}, i = 0,1 \ldots \tag{14}$$

$$\tilde{y} = arg\max_i p_i \tag{15}$$

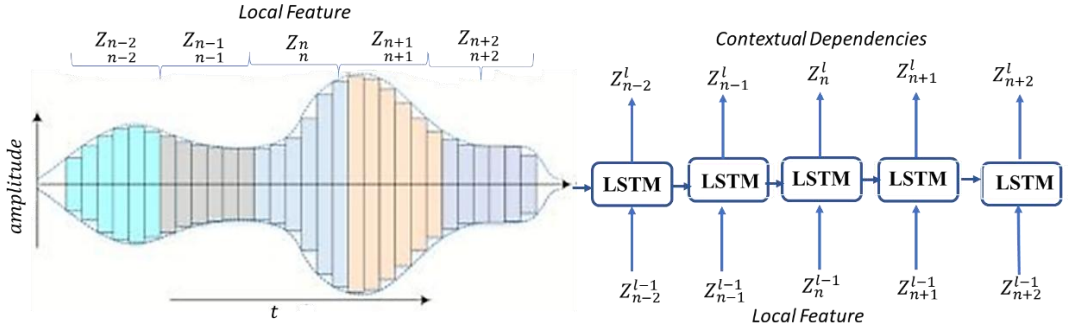


Figure 6. 1D CNN+LSTM

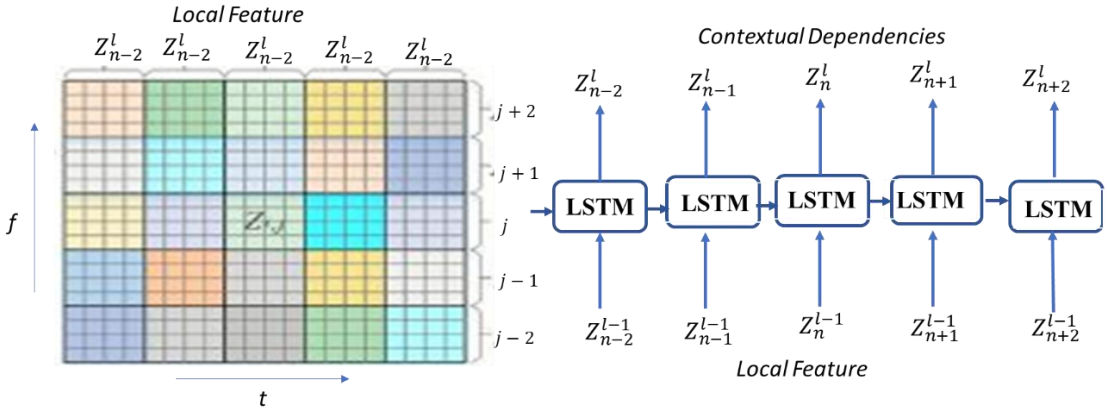


Figure 7. 2D CNN+LSTM

To reduce the hazards of overfitting and underfitting in trials, numerous innovative solutions have been proposed. Overfitting happens when a model memorizes data rather than improving prediction, which is usually caused by a complex or overtrained deep network. Methods like cross-validation, regularization, BN, and early pausing are employed to overcome this. The experimental data used in our investigation was split at random into two sets: training (80%) and testing (20%) for SER with good accuracy and generalization. Only accurate models are recorded to avoid overfitting, which occurs when extensive training leads to poor performance on new data. Early stopping checks model performance on a validation dataset and stops training when the validation error no longer improves.

## 5. EXPERIMENTAL RESULTS

The SER_DLFNet, a fusion-based DL network for SER, was designed, trained, and evaluated with RAVDESS. The model's performance was evaluated using objective metrics such as average accuracy, precision, and recall. The number of training epochs in our proposed network was chosen iteratively by experimenting with different numbers and comparing performance on a validation dataset. After evaluating various epoch settings, it was discovered that the model with 500 epochs achieved the best balance of underfitting and overfitting, resulting in optimal performance on the validation dataset. The hybrid CNN+LSTM networks were trained and evaluated on the RAVDESS emotional speech dataset, which contains audio clips sampled at a rate of 16 kHz and with a fixed duration of eight seconds. If an audio clip exceeds eight seconds, it is segmented down, while shorter clips are padded to meet the eight-second length. At the 16 kHz sample rate, a 128,000-bit vector describes each audio clip.

The FFT window length is configured to 2048, with a hop length of 512 during the calculation of the log-Mel spectrogram. This process produces a log-Mel spectrogram comprising 251 frames and 128 Mel frequency bins. The visualization of the log-Mel spectrogram can be presented as either a grid or a sequence for analysis. In our experiments, the 128×251 matrices served as input to the 2D Hybrid network, enabling the 2D-CNN+LSTM network to learn high-level features from these image-like patches.

### 5.1. 1D-CNN

The proposed 1D CNN model for SER consists of six convolutional layers, each utilizing a Kernel size of 1×5, followed by a single fully connected layer. This design incorporates zero padding and ReLu activation functions to preserve information along the edges. Additionally, the model employs maximum pooling with a size of 1×8, complemented by a dropout rate of 0.1 and BN. Through training this network model for 500 epochs, an impressive average accuracy of 85.47% was achieved. The corresponding confusion matrix for the network model, demonstrating an accuracy matching the aforementioned percentage, is presented in Figure 8. These parameters collectively contribute to the effectiveness of the proposed 1D CNN model in recognizing emotions from speech.

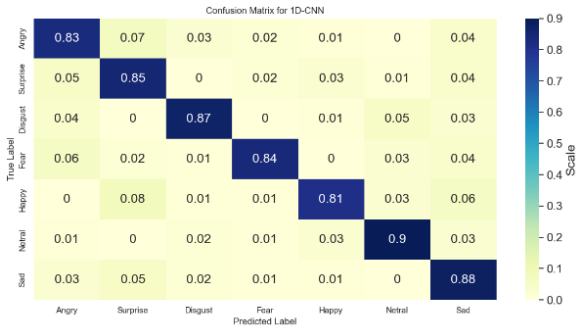


Figure 8. Confusion matrix of 1D-CNN

### 5.2. 2D-CNN

In the realm of 1D CNN, the convolution operation is limited to a single direction, whereas the 2D CNN's convolution Kernel operates in two dimensions, opening possibilities for applications in studying time series data. The experiment was structured with six convolution layers, each featuring a 3×3 filter size, zero padding to preserve edge information, and ReLU activation. The max-pooling procedure, executed on a 2×2 matrix, incorporated a dropout of 0.1 and BN. Training the network model for 500 epochs yielded an impressive accuracy of 88.36%. The accompanying confusion matrix, illustrated in Figure 9, provides a visual representation of the network model's enhanced accuracy. This configuration demonstrates the efficacy of the 2D CNN model in handling time series data and achieving improved performance in emotion recognition tasks.

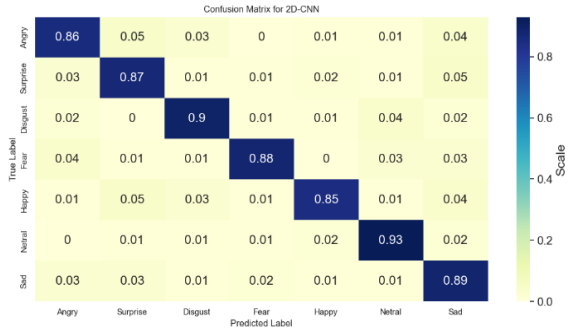


Figure 9. Confusion matrix of 2D-CNN

At last, LFLBs and LSTM were combined to create hybrid DL network models. Every LFLB included a single stride, a 3×3 filter size, and no padding. Tables 1 and 2 illustrate strutures of LFLB and LSTM, that the first and second LFLBs each had 64 convolution Kernels, the third and fourth LFLBs had 128 convolution Kernels, and the fifth and sixth LFLBs had 256. Following training and testing, the suggested network models' effectiveness was assessed using objective metrics on the RAVDESS database across a range of epoch counts.

Table 1. Structure of 4 LFLB+LSTM

| Block name | | Output dimension | Kernel size | Stride |
|---|---|---|---|---|
| LFFB1 | 1C1 | L×64 | 3×3 | 1×1 |
| | 1P1 | L/4×64 | 4×4 | 4×4 |
| LFFB2 | 1C2 | L/4×64 | 3×3 | 1×1 |
| | 1P2 | L/16×64 | 4×4 | 4×4 |
| LFFB3 | 1C3 | L/16×128 | 3×3 | 1×1 |
| | 1P3 | L/64×128 | 4×4 | 4×4 |
| LFFB4 | 1C4 | L/64×128 | 3×3 | 1×1 |
| | 1P4 | L/256×128 | 4×4 | 4×4 |
| L | | - | 256×1 | - |
| F | | - | K | - |

Table 2. Structure of 6 LFLB+LSTM

| Block name | | Output dimension | Kernel size | Stride |
|---|---|---|---|---|
| LFFB1 | 1C1 | L×64 | 3×3 | 1×1 |
| | 1P1 | L/4×64 | 4×4 | 4×4 |
| LFFB2 | 1C2 | L/4×64 | 3×3 | 1×1 |
| | 1P2 | L/16×64 | 4×4 | 4×4 |
| LFFB3 | 1C3 | L/16×128 | 3×3 | 1×1 |
| | 1P3 | L/64×128 | 4×4 | 4×4 |
| LFFB4 | 1C4 | L/64×128 | 3×3 | 1×1 |
| | 1P4 | L/256×128 | 4×4 | 4×4 |
| LFFB5 | 1C5 | L/256×128 | 3×3 | 1×1 |
| | 1P5 | L/256×128 | 4×4 | 4×4 |
| LFFB6 | 1C6 | L/512×256 | 3×3 | 1×1 |
| | 1P6 | L/512×256 | 4×4 | 4×4 |
| L | | - | 512×1 | - |
| F | | - | K | - |

### 5.3. 1D-CNN+LSTM

Initially, a 1D-CNN+LSTM network model was introduced with 4LFLB+LSTM layers, 64 fully connected neurons, a batch size of 32, and a learning rate of 0.0001 utilizing the stochastic gradient descent optimizer. The model was trained for several epochs, yielding accuracies of 77.35%, 78.65%, 90.94%, and 71.80% for 50, 100, 300, and 500 epochs, respectively. However, due to underfitting, the model's accuracy varied over epochs, resulting in frequent misclassification of some classes and contributing to the network model's overall low accuracy. To improve the model's performance, we did additional training on the most complex emotional classes, such as cases where happy was commonly misclassified as neutral and fear as anger. This involved fine-tuning hyperparameters to achieve the optimal balance between underfitting and overfitting, leading to improved performance on the validation dataset.

The network parameters were modified to include 6LFLB+LSTM layers, a batch size of 32, and a learning rate of 0.001. The model was then retrained for 50, 100, 300, and 500 epochs, yielding accuracy rates of 79.96%, 81.53%, 91.51%, and 80.25%, respectively. Figure 10 shows the accuracy vs the number of epochs for the 1D-CNN+LSTM, which includes two distinct networks. Furthermore, Figure 11 depicts the confusion matrix for improved accuracy of the 1D-CNN+LSTM network model after 300 epochs, with an average accuracy of 91.51%, precision of 91.62%, and recall of 91.50%.

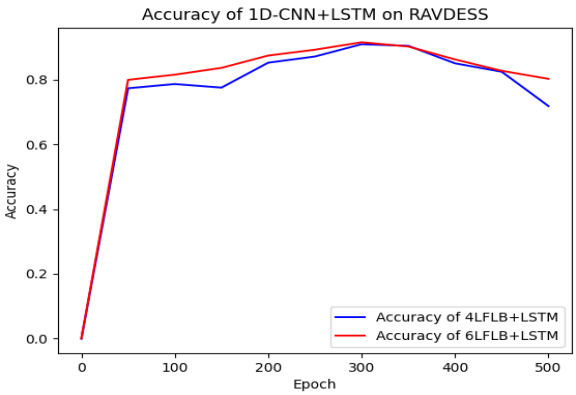


Figure 10. 1D-CNN and LSTM: accuracy vs epochs

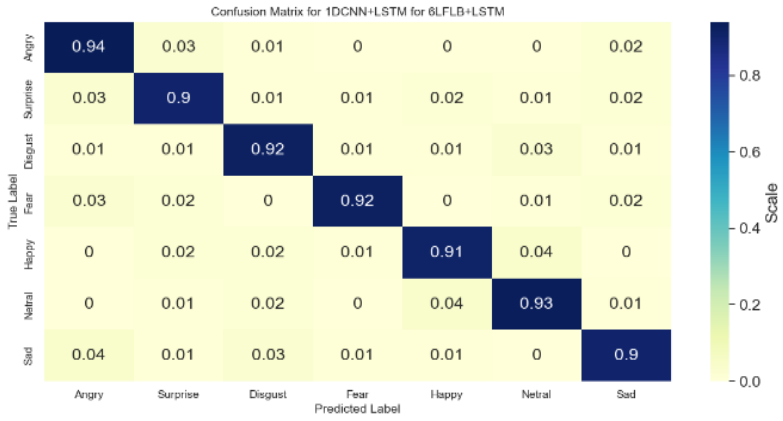


Figure 11. Confusion matrix of 1D-CNN and LSTM for 300 epochs with 6 LFLB and LSTM

### 5.4. 2D-CNN+LSTM

In a manner similar to the 1D-CNN+LSTM network model, the 2D-CNN+LSTM network model also underwent an initial configuration. A network with 4LFLB+LSTM layers was proposed, featuring 64 fully connected neurons, a batch size of 32, a learning rate of 0.0001, and a momentum of 0.9. Stochastic gradient descent optimizer was utilized, and the model was trained for different epoch counts, yielding accuracies of 79.50%, 82.45%, 93.64%, and 85.72% for 50, 100, 300, and 500 epochs, respectively.

To improve the model's performance, we modified the architecture to 6LFLB+LSTM and trained it again on the most complicated emotional classes. This involved adding two additional LFLBs while keeping the same hyperparameters as the 2D 4LFLB+LSTM configuration. The corresponding accuracies for 50, 100, 300, and 500 epochs were 80.85%, 84.76%, 95.52%, and 87.95%. Table 3 summarizes the proposed model accuracies across various epoch counts. Figure 12 shows the accuracy versus the number of epochs for the 1D-CNN+LSTM, which uses two distinct networks. Furthermore, Figure 13 depicts the confusion matrix for the enhanced accuracy network model after 300 epochs, with an average accuracy of 95.52%, precision of 95.61%, and recall of 95.51%.

Table 3. CNN and LSTM model accuracies vs Number of Epochs

| Number of epochs | 1D-CNN and LSTM | | 2D-CNN and LSTM | |
|---|---|---|---|---|
| | Categorical accuracy of 4LFLB+LSTM network (%) | Categorical accuracy of 6LFLB+LSTM network (%) | Categorical accuracy of 4LFLB+LSTM network (%) | Categorical accuracy of 6LFLB+LSTM network (%) |
| 50 | 77.35 | 79.96 | 79.50 | 80.85 |
| 100 | 78.65 | 81.53 | 82.45 | 84.76 |
| 300 | 90.94 | 91.51 | 93.64 | 95.52 |
| 500 | 71.80 | 80.25 | 85.72 | 87.95 |

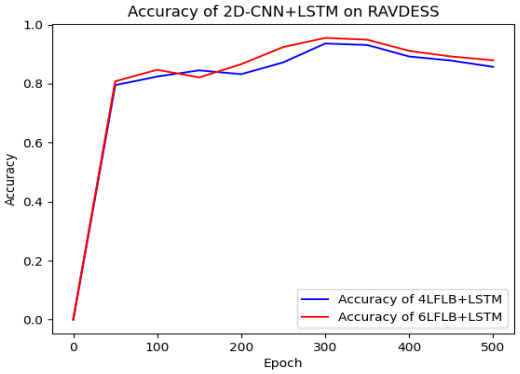


Figure 12. 2D-CNN and LSTM: accuracy vs epochs

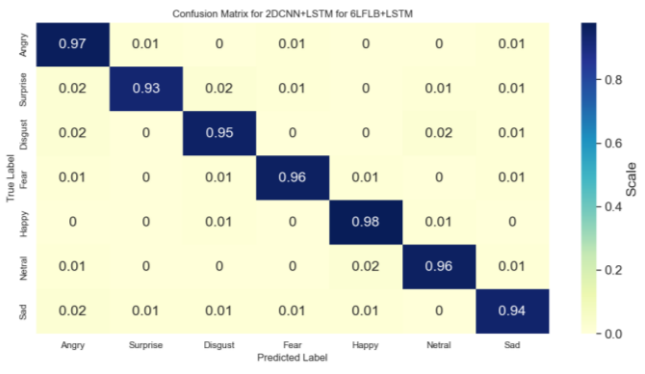


Figure 13. Confusion matrix of 2D-CNN and LSTM for 300 epochs with 6 LFLB and LSTM

## 6. RESULT ANALYSIS

Initially, each model was given a unique architecture and set of hyperparameters before being trained over a number of epochs. Accuracy levels were documented to evaluate each network model's first performance. Subsequently, to improve the performance of each network model, the architecture and hyperparameters were fine-tuned, with an emphasis on reducing confusion in specific classes. This adjustment resulted in significant improvements in the accuracy of the network models. Table 4 provides a succinct overview of the proposed network models, including their best accuracies. The findings demonstrate that the 2D hybrid network outperforms existing established methods in terms of accuracy, precision, and recall.

Table 4. Different proposed network models accuracies for proposed system

| Objective metrics\network model | 1-D CNN | 2-D CNN | 1-D CNN and LSTM | 2-D CNN and LSTM |
|---|---|---|---|---|
| Accuracy | 85.47 | 88.36 | 91.51 | 95.52 |
| Precision | 85.79 | 88.46 | 91.62 | 95.61 |
| Recall | 85.42 | 88.28 | 91.49 | 95.50 |

In this study, the proposed model was compared with existing DL algorithms, as presented in Table 5. In our proposed model, the use of the ELU activation function, rather than ReLU, has played a critical role in achieving superior accuracy compared to earlier studies by Zeng *et al*. [26] who employed a single CNN, and by Jalal *et al.* [27] and Mustaqeem *et al*. [22], who used a hybrid network model (CNN+LSTM) with ReLU activation function and only three convolution layers in their models with two LSTM layers [3]. ELU has a significant benefit over ReLU in that it allows for the generation of negative outputs, which allows the neural network to comprehend more complex and expressive characteristics. The ELU activation function produces smoother outputs than ReLU and addresses the issue of dying ReLU, in which certain ReLU neurons become inactive and emit zero outputs during training, resulting in a dormant neuron that contributes little to the network's learning process [16].

Table 5. Comparison of proposed versus other model performances on RAVDESS

| Method | Accuracy (%) | Precision (%) | Recall (%) |
|---|---|---|---|
| Zeng *et al*. [26] | 64.48 | - | - |
| Jala *et al.* [27] | 69.40 | - | - |
| Mustaqeem *et al*. [22] | 91.14 | - | - |
| Proposed model | 95.52 | 95.61 | 95.50 |

## 7. CONCLUSION

This chapter examines 1D CNN, 2D CNN, 1D CNN and LSTM, and 2D CNN and LSTM DL network models for SER. The goal is to understand local correlations and global context using both raw audio recordings and speech signal spectrograms. To capture local features, a LFLB is used and an LSTM layer handles the global features, which consists of context dependent information. Initially, each model was proposed with a distinct architecture and set of hyperparameters, and it was trained over a variety of epoch counts. The resulting accuracies were measured to assess the network models' performance. To improve the efficacy of each network model, the architecture and hyperparameters were fine-tuned, with a focus on addressing the issues presented by the most confusing classes. This adjustment resulted in significant improvements in the accuracy of the network models. It is critical to understand that the appropriate architecture and hyperparameters for a DL network model might differ depending on the task and dataset. As a result, testing with various configurations is necessary to determine the most effective one, taking into account aspects such as the quality and quantity of training data.

## REFERENCES

[1] Z. T. Liu, Q. Xie, M. Wu, W. H. Cao, Y. Mei, and J. W. Mao, "Speech emotion recognition based on an improved brain emotion learning model," *Neurocomputing*, vol. 309, pp. 145–156, Oct. 2018, doi: 10.1016/j.neucom.2018.05.005.

[2] D. Liu, Z. Wang, L. Wang, and L. Chen, "Multi-Modal Fusion Emotion Recognition Method of Speech Expression Based on Deep Learning," *Frontiers in Neurorobotics*, vol. 15, Jul. 2021, doi: 10.3389/fnbot.2021.697634.

[3] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, Jan. 2019, doi: 10.1016/j.bspc.2018.08.035.

[4] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019, doi: 10.1109/ACCESS.2019.2936124.

[5] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods,

supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, Jan. 2020, doi: 10.1016/j.specom.2019.12.001.
[6] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artificial Intelligence Review*, vol. 53, no. 8, pp. 5455–5516, Dec. 2020, doi: 10.1007/s10462-020-09825-6.
[7] C. A. Kumar and K. A. Sheela, "Emotion Recognition from Speech Biometric System Using Machine Learning Algorithms," in *Lecture Notes in Electrical Engineering*, 2021, pp. 63–73, doi: 10.1007/978-981-33-4058-9_6.
[8] C. A. Kumar and K. A. Sheela, "Emotion Recognition from Facial Biometric System Using Deep Convolution Neural Network (D-CNN)," in *Lecture Notes in Mechanical Engineering*, 2021, pp. 381–391, doi: 10.1007/978-981-15-8025-3_37.
[9] C. A. Kumar and K. A. Sheela, "Real-Time Emotional Analysis from A Live Webcam Using Deep Learning," in *2022 3rd International Conference for Emerging Technology, INCET 2022*, IEEE, May 2022, pp. 1–5, doi: 10.1109/INCET54531.2022.9824894.
[10] O. W. Chuan, N. F. Ab Aziz, Z. M. Yasin, N. A. Salim, and N. A. Wahab, "Fault classification in smart distribution network using support vector machine," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 18, no. 3, pp. 1148–1155, Jun. 2020, doi: 10.11591/ijeecs.v18.i3.pp1148-1155.
[11] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, Mar. 2011, doi: 10.1016/j.patcog.2010.09.020.
[12] A. D. Dileep and C. C. Sekhar, "GMM-based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 8, pp. 1421–1432, Aug. 2014, doi: 10.1109/TNNLS.2013.2293512.
[13] O. Abdel-Hamid, A. R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 10, pp. 1533–1545, Oct. 2014, doi: 10.1109/TASLP.2014.2339736.
[14] Y. Tang, "Deep Learning using Linear Support Vector Machines," *arXiv*, Jun. 2013, doi: 10.48550/arXiv.1306.0239
[15] K. S. Rao, S. G. Koolagudi, and R. R. Vempada, "Emotion recognition from speech using global and local prosodic features," *International Journal of Speech Technology*, vol. 16, no. 2, pp. 143–160, Jun. 2013, doi: 10.1007/s10772-012-9172-2.
[16] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, IEEE, Jul. 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.
[17] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in Neural Information Processing Systems*, vol. 4, no. January, pp. 3104–3112, Sep. 2014.
[18] M. S. Hossain and G. Muhammad, "An Audio-Visual Emotion Recognition System Using Deep Learning Fusion for a Cognitive Wireless Framework," *IEEE Wireless Communications*, vol. 26, no. 3, pp. 62–68, Jun. 2019, doi: 10.1109/MWC.2019.1800419.
[19] L. Schoneveld, A. Othmani, and H. Abdelkawy, "Leveraging recent advances in deep learning for audio-Visual emotion recognition," *Pattern Recognition Letters*, vol. 146, pp. 1–7, Jun. 2021, doi: 10.1016/j.patrec.2021.03.007.
[20] S. Ntalampiras, "Speech emotion recognition via learning analogies," *Pattern Recognition Letters*, vol. 144, pp. 21–26, Apr. 2021, doi: 10.1016/j.patrec.2021.01.018.
[21] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial-Temporal Recurrent Neural Network for Emotion Recognition," *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 939–947, Mar. 2019, doi: 10.1109/TCYB.2017.2788081.
[22] Mustaqeem, M. Sajjad, and S. Kwon, "Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM," *IEEE Access*, vol. 8, pp. 79861–79875, 2020, doi: 10.1109/ACCESS.2020.2990405.
[23] G. Du, Z. Wang, B. Gao, S. Mumtaz, K. M. Abualnaja, and C. Du, "A Convolution Bidirectional Long Short-Term Memory Neural Network for Driver Emotion Recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4570–4578, Jul. 2021, doi: 10.1109/TITS.2020.3007357.
[24] A. Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, Mar. 2020, doi: 10.1016/j.physd.2019.132306.
[25] Y. Eom and J. Bang, "Speech Emotion Recognition Using 2D-CNN with Mel-Frequency Cepstrum Coefficients," *Journal of Information and Communication Convergence Engineering*, vol. 19, no. 3, pp. 148–154, Sep. 2021, doi: 10.6109/jicce.2021.19.3.148.
[26] Y. Zeng, H. Mao, D. Peng, and Z. Yi, "Spectrogram based multi-task audio classification," *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 3705–3722, Feb. 2019, doi: 10.1007/s11042-017-5539-3.
[27] Md. A. Jalal, R. Milner, and T. Hain, "Empirical interpretation of speech emotion perception with attention based model for speech emotion Recognition," In: *Proceedings of Interspeech 2020*, Shanghai, China (Online). International Speech Communication Association (ISCA), 25-29 Oct 2020, pp. 4113-4117, doi: 10.21437/Interspeech.2020-3007.

## BIOGRAPHIES OF AUTHORS

**Chevella Anil Kumar** is an Assistant Professor in the Department of ECE ,VNR Vignana Jyothi Institute of Engineering and Technology in Hyderabad, Telangana, India. He was awarded a Doctor of Philosophy (Ph.D.) from JNT University Hyderabad. His research focuses on image, speech, and video signal processing, neural networks, machine learning, and deep learning. He can be contacted at email: chevellaanilkumar@gmail.com and anilkumar_chevella@vnrvjiet.in.

**Vumanthala Sagar Reddy** is an Assistant Professor, Department of ECE, VNR Vignana Jyothi Institute of Engineering and Technology in Hyderabad, Telangana, India. He was awarded a Doctor of Philosophy (Ph.D.) from Kaktiya University Warangal. His research focuses on image, speech, VLSI signal processing, and machine learning. He can be contacted at email: vsagarreddy1990@gmail.com and sagar_v@vnrvjiet.in.

**Ambati Pravallika** is an Assistant Professor, Department of ECE, VNR Vignana Jyothi Institute of Engineering and Technology in Hyderabad, Telangana, India. Pursuing Ph.D at NIT Warangal in the Domain of Image and video Processing. Her research area is 3D object recognition and tracking using deep learning. She can be contacted at email: pravallika_a@vnrvjiet.in.

**Rao Y. Chalapathi** Associate Professor, Department of ECE, VNR Vignana Jyothi Institute of Engineering and Technology in Hyderabad, Telangana, India. He had more than 20 years experience in teaching and his research focuses on signal processing (image and video), communication using machine learning, and deep learning. He can be contacted at email: chalu.8421@gmail.com.

**Neelam Syamala** Assistant Professor, Department of ECE, VNR Vignana Jyothi Institute of Engineering and Technology in Hyderabad, Telangana, India. Pursuing Ph.D. at Vellore Institute of Technology, Vellore, Chennai, Tamilnadu, India. Her research interests are in the domain of signal processing and communication. he can be contacted at email: syamala_n@vnrvjiet.in.