2025 International Conference on Electrical, Computer and Communication Engineering (ECCE),13-15 February 2025, CUET, Chattogram-4349, Bangladesh

Revamping EmoTech: Leveraging Dilated Convolution for Cutting-Edge Robust Speech Emotion Recognition

Nur Alam, Nuzhat Mobassara, Nursadul Mamun
Robust Speech Processing Laboratory (RSPL)
Department of Electronics and Telecommunication Engineering
Chittagong University of Engineering and Technology, Chittagong
u1908046@student.cuet.ac.bd, u1908006@student.cuet.ac.bd, nursad.mamun@cuet.ac.bd

***Abstract*—Emotion recognition plays a crucial role in enhancing human-computer interaction by enabling systems to respond naturally to users' emotions. Traditional approaches to speech emotion recognition (SER) often relied on manual feature extraction and conventional machine learning models, which struggled to capture complex emotional patterns and long-range dependencies in speech. This paper presents a robust SER framework that addresses these limitations by leveraging advanced feature extraction techniques and innovative architectural elements. The proposed model integrates dilated convolutional layers with Bidirectional Long Short-Term Memory Networks (BiLSTMs) to effectively capture long-range dependencies and intricate emotional patterns in speech signals. Using Mel Frequency Cepstral Coefficients (MFCCs) as the primary features, the network processes data through two parallel pathways: one employing a BiLSTM network for sequence modeling and the other utilizing dilated convolutions to extract temporal spectral features. To address data imbalance and enhance model robustness, data augmentation techniques such as pitch shifting and time-stretching are applied. The fused features are then fed into fully connected layers to classify emotions into 10 classes for IEMOCAP and 6 classes for CREMA-D. Experimental results on the IEMOCAP and CREMA-D datasets demonstrate the model's effectiveness, achieving test accuracies of 87% and 86%, respectively. Incorporating dilated convolutions and data augmentation significantly improves the recognition of subtle emotional cues, offering a reliable and scalable solution for speech emotion recognition tasks.**

***Index Terms*—Speech Emotion Recognition, Dilated Convolution, BiLSTM, Data Augmentation, MFCC, Robustness**

I. Introduction

Speech Emotion Recognition (SER) has emerged as a critical domain within human-computer interaction (HCI), leveraging advancements in artificial intelligence and signal processing. Emotions are pivotal in human communication, influencing decision-making, interpersonal interactions, and overall cognitive functioning. By bridging the gap between human emotional intelligence and machine understanding, SER enables systems to detect and interpret emotional states from speech signals, fostering more intuitive and empathetic interactions.

The significance of SER extends across diverse applications, from healthcare to security. In healthcare, it facilitates early detection of mental health issues like depression and anxiety by analyzing vocal cues. In call centers, SER enhances customer experience through real-time insights into emotional states, enabling adaptive responses. It also supports personalized learning in education via emotion-aware virtual tutors and aids in stress or deception detection in security and law enforcement. Despite its transformative potential, achieving robust emotion recognition remains challenging due to variability in speech patterns, cultural differences, and real-world noise. Recent studies, such as [6], [12], [14], address these challenges, proposing innovative solutions to advance the field of SER.

In recent years, numerous studies have explored auditory features and methodologies to enhance speech emotion recognition (SER) systems [2], [3], [9], [13], [15]. Early machine learning approaches relied on handcrafted features, such as Mel Frequency Cepstral Coefficients (MFCC), for training models [4]. However, these traditional models, including support vector machines and Gaussian mixture models, often struggled to generalize across diverse speakers and languages, as highlighted by Ayadi et al. [5].

The advent of deep learning introduced powerful alternatives, with Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) becoming the standard for SER. CNNs excel at extracting spectral features from audio signals, while RNNs, particularly Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) networks, capture temporal dependencies effectively [1], [16]. Hybrid architectures, such as those proposed by Satt et al., combine CNNs for spectral feature extraction with RNNs for temporal analysis, achieving improved performance [10]. Expanding on this, Avro et al. introduced the multimodal architecture Emotech, which integrates audio and text modalities for emotion recognition. Emotech employs CNN and BiLSTM layers within its audio and text blocks, enabling the model to capture both local features and temporal dependencies, effectively disentangling complex emotions even in challenging real-world conditions.

While Emotech demonstrates strong performance when combining text data with speech signals, its effectiveness decreases when limited to speech-only data. This limitation arises

from the restricted receptive field of convolutional layers, which hampers the model's ability to differentiate structurally dependent emotions, such as neutral versus happy or excited states. Additionally, synchronizing textual data with speech signals can be challenging under real-world conditions.

To address these challenges, this study proposes a novel network— a uni-modal architecture for SER using speech signals. The proposed model incorporates dilation layers into CNN-RNN architectures, enhancing the network's ability to process complex emotional cues. Dilation layers expand the receptive field of convolutional layers without increasing the number of parameters, enabling the model to capture both fine-grained local details and broader contextual information. By integrating dilation layers with CNNs and RNNs, the architecture effectively combines local feature extraction, temporal dependencies, and multiscale feature representation. This approach is particularly beneficial in SER, where emotional cues are distributed across multiple time scales. Additionally, it enhances the model's sensitivity to both subtle and prominent emotional features, improving accuracy and robustness across diverse datasets and noisy environments.

The key contributions of this study are as follows:

1) A novel speech-based SER architecture that incorporates dilation layers to effectively expand the receptive field for improved emotion recognition in speech signals.
2) The integration of BiLSTM and CNN layers within the network to capture both temporal dependencies and local features, enhancing the model's ability to interpret complex emotional cues.
3) A comprehensive evaluation of the proposed model on diverse speech emotion datasets, demonstrating its superior performance compared to existing methods.

The paper is organized as follows: Section 2 describes the individual parameters of the proposed network used in this research. Section 3 presents the experimental results, followed by the conclusion in Section 4.

II. Methodology

In this work, a new SER network was introduced based on audio features alone. The proposed method fuses deep learning techniques to effectively identify emotional states extracted from speech signals from two most commonly used datasets: CREMA-D and IEMOCAP. The primary issue of this investigation is to construct a robust SER model that can address different speech emotions and achieve excellent accuracy even when faced with class imbalances.

The process, therefore, includes several steps: data augmentation, feature extraction, and model development, all of which go in the direction of improving recognition efficacy in different emotional conditions.

A. Dataset

The CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) contains approximately 7,000 audio samples from 91 actors portraying six emotions: Anger, Disgust, Fear, Happiness, Neutral, and Sadness. Each sample includes an audio file and an emotion label aligned with the actor's vocal expression. For this study, a subset of 2,930 samples was selected to ensure a balance between dataset size and emotional diversity.

The IEMOCAP (Interactive Emotional Dyadic Motion Capture) dataset consists of approximately 12 hours of audio-visual recordings from 10 actors (5 male and 5 female) expressing a broader range of 10 emotions: Neutral, Sadness, Excitement, Anger, Surprise, Happiness, Fear, Disturbance, Frustration, and others. A representative subset of 2,779 samples was selected from Sessions 1 and 2. After further balancing, a final total of 7,460 samples was curated for analysis.

These curated datasets collectively provide a diverse and robust foundation for evaluating emotion recognition models, ensuring a comprehensive analysis of various emotional states.

B. Feature Extraction

To extract meaningful information from the audio signal, both MFCCs and Mel-spectrograms were utilized, capturing critical elements such as pitch, tone, and patterns essential for emotion recognition. For MFCC extraction, 13 coefficients were calculated for each audio sample to represent the spectral envelope of speech. These coefficients were averaged along the time axis, producing a fixed-length feature vector. Similarly, the Mel-spectrogram was derived to represent speech features across time and frequency, leveraging the Mel scale's alignment with human auditory perception. Averaging over time resulted in 128-dimensional feature vectors. By combining the MFCC and Mel-spectrogram features, a comprehensive 141-dimensional feature vector was generated for each sample, ensuring a robust representation of both time- and frequency-based information. Mel-frequency cepstral coefficients (MFCCs) and Mel-spectrograms have been chosen to extract basic speech features associated with emotional expression. MFCCs capture the spectral envelope and represent those changes in pitch, tone, and loudness vital for emotion recognition. In contrast, Mel-spectrograms represent the time-frequency of a sound, thus showing the dynamic development of emotions. Such characteristics relate to human auditory perception and increase the capability of modeling to identify emotional cues from a person's voice. Alternatives like Linear Predictive Coding (LPC), prosodic features, and raw waveform-based methods like WaveNet provide promising alternatives; however they might require more resources or neglect subtle emotional cues. CNNs, on the other hand, can directly learn features from spectrograms.

C. Data Augmentation

The datasets used in this study exhibited notable class imbalances, with certain emotions being underrepresented. In the CREMA-D dataset, emotions such as Neutral had fewer samples, while in the IEMOCAP dataset, emotions like Happiness, Surprise, and Fear were underrepresented. To address this issue, random over-sampling was applied by duplicating samples from minority classes to achieve equal

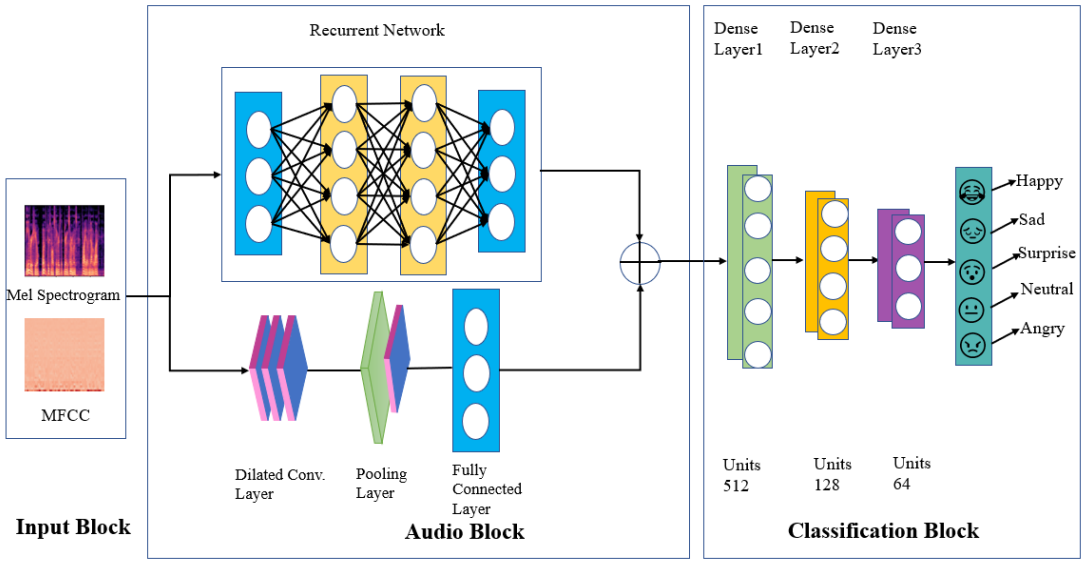


Fig. 1: Overall block diagram of the proposed SER network

class representation. This process increased the sample count in CREMA-D from 2,930 to 3,072, ensuring a balanced representation across six emotions. Similarly, in IEMOCAP, the sample count increased from 2,779 to 7,460, achieving a uniform distribution for all emotions across Sessions 1 and 2. Figures 2a and 2b illustrate the emotion distributions before and after balancing for both datasets. To further enhance the robustness of the model and mitigate overfitting, data augmentation techniques were applied. Two additional versions of each audio sample were generated using pitch shifting and time stretching. Pitch shifting adjusted the pitch by two semitones, introducing variability in vocal tone, while time stretching increased the audio length by 1.2 times its original speed, simulating variations in speaking pace. As a result of these augmentations, the number of samples in CREMA-D increased from 3,072 to 12,288, while in IEMOCAP, the sample count increased from 7,460 to 22,380. These additional variations enriched the datasets by introducing greater diversity, ultimately improving the reliability and generalization capability of the trained models.

D. Dilated Convolution

Dilated convolution, also known as atrous convolution, is a convolution operation with gaps (dilation) between kernel elements that increases the receptive field without increasing the number of parameters or computational cost. It is an operation in which the kernel slides over the input, not necessarily with every element in standard convolution; through dilation, the model could potentially capture long-range dependencies.

Mathematically, dilated convolution is defined as:

$$y[i] = \sum_{k=0}^{K-1} x[i + k \cdot d] \cdot w[k],$$

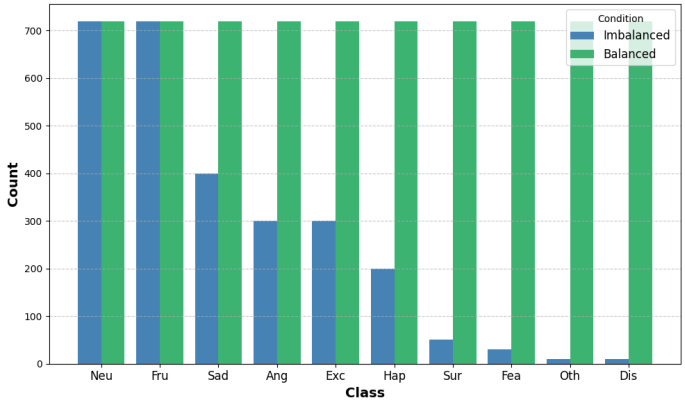
where $y[i]$ represents the output at position i, $x[i]$ is the input signal, $w[k]$ is the convolutional kernel of size K, and d is the dilation rate. A dilation rate $d = 1$ reduces this operation to standard convolution.

The dilated convolutions with a dilation rate of $(2, 1)$ are proposed to be used in this model to extract the native temporal structures of the audio signals effectively. The dilation rate of 1 is supposed to permit capturing short-term, fine-grained features like the variations of pitch and tone that are vital in characterizing immediate emotional change, while the dilation rate of 2 will let the model capture longer-term dependencies, including sustained emotions and speech transitions. This combination enhances the model's ability to recognize rapid emotional changes and more significant emotional trends. Overall, these dilation rates simplify feature extraction, translating to improved speech emotion recognition. Unlike in standard convolution, dilated convolution learns features over much larger time windows, which would further increase the model's capability to recognize the emotions expressed through speech.

E. Network Architecture

The proposed model architecture, illustrated in Fig. 1, processes Mel Frequency Cepstral Coefficients (MFCCs) as the primary audio features. The input audio signals are resampled to 16 kHz, and 13 cepstral coefficients are extracted. To ensure uniform input dimensions, zero-padding is applied to the audio features. The MFCCs are processed through two parallel paths: a recurrent network using Bidirectional Long Short-Term Memory (BiLSTM) layers and a convolutional network employing 2D convolutions with dilated kernels.

The BiLSTM branch consists of two layers, each with 64 hidden units and the tanh activation function, generating an output tensor with a shape of (128,). The convolutional path



(a)

(b)

Fig. 2: Emotion distribution (a) IEMOCAP dataset, (b) CREMA-D dataset.

utilizes three Conv2D layers with dilation rates of (2,1) and kernel sizes of (3,3), featuring progressively increasing kernel counts of 32, 64, and 128. Each convolutional layer is followed by Batch Normalization to stabilize training and MaxPooling with a window size of (2,2) to reduce spatial dimensions while preserving critical features. The output of the convolutional block is reshaped into a tensor suitable for further processing.

The reshaped convolutional output is passed through a dense network composed of two fully connected layers with 512 and 128 units. Both layers incorporate a dropout rate of 0.2 to mitigate overfitting, producing an output vector with a shape of (128,). This vector is concatenated with the (128,) output from the BiLSTM branch, forming a combined feature representation with a shape of (256,). The concatenated feature vector is fed into a classification block comprising three fully connected layers with 512, 128, and 64 units, respectively, which map the features to predefined emotion classes such as Happy, Sad, Surprise, Neutral, Angry, and others specified in the datasets.

The incorporation of dilated convolutions enables the model to capture long-term dependencies in the audio signals, enhancing its ability to recognize intricate emotional patterns. The model has 313K parameters, reflecting its complex architecture. Training was conducted on Google Colab using a T4 GPU and on Kaggle using a P100 GPU, both of which significantly accelerated the training process by efficiently handling the model's parameter space.

III. Results

To evaluate the performance of the proposed model, the network was tested on two datasets: CREMA-D and IEMOCAP. The evaluation was conducted with and without data augmentation across different classifiers. Additionally, the performance of the proposed model was compared against several baseline networks.

A. Performance Analysis on Proposed Model

The performance of the proposed model was evaluated on the CREMA-D dataset, revealing insightful trends. Without applying data augmentation, the baseline model achieved an average validation accuracy of 49%. Similarly, the model using dilated convolutions demonstrated comparable performance, also achieving 49% precision (Table I). This outcome suggests that in the absence of sufficient data diversity, the benefits of dilated convolutions are minimal.

TABLE I: Performance Comparison with and without Data Augmentation on IEMOCAP and CREMA-D Datasets

Dataset	Model	Before Aug.(%)	After Aug.(%)
IEMOCAP	EmoTech [3]	70	71
	Proposed	**84**	**87**
CREMA-D	EmoTech [3]	49	83
	Proposed	**49**	**86**

However, a dramatic improvement was observed when data augmentation techniques were introduced. The baseline EmoTech model [3] showed a substantial increase in accuracy, improving from 49% to 83%. The proposed model outperformed all others, achieving the highest accuracy of 86% (Table I). These results highlight the critical role of data augmentation in enhancing the model's generalizability and demonstrate the ability of dilated convolutions to capture intricate temporal patterns, resulting in improved emotion recognition.

The confusion matrix for the CREMA-D dataset, presented in Figure 3, offers a detailed analysis of the model's classification performance. High values along the diagonal confirm the model's ability to predict the correct emotions accurately. Meanwhile, off-diagonal values indicate misclassifications, especially between the classes of fear and sadness (166 cases) and disgust and sadness (125 cases), likely due to overlapping acoustic features such as low-pitched and monotonic tone. The subtle cues of fear and the prosodic resemblance of disgust to sadness cause confusion. A possible reason for the neutral misclassification of fear (59 cases) could be the low emotional intensity. These errors may stem from feature

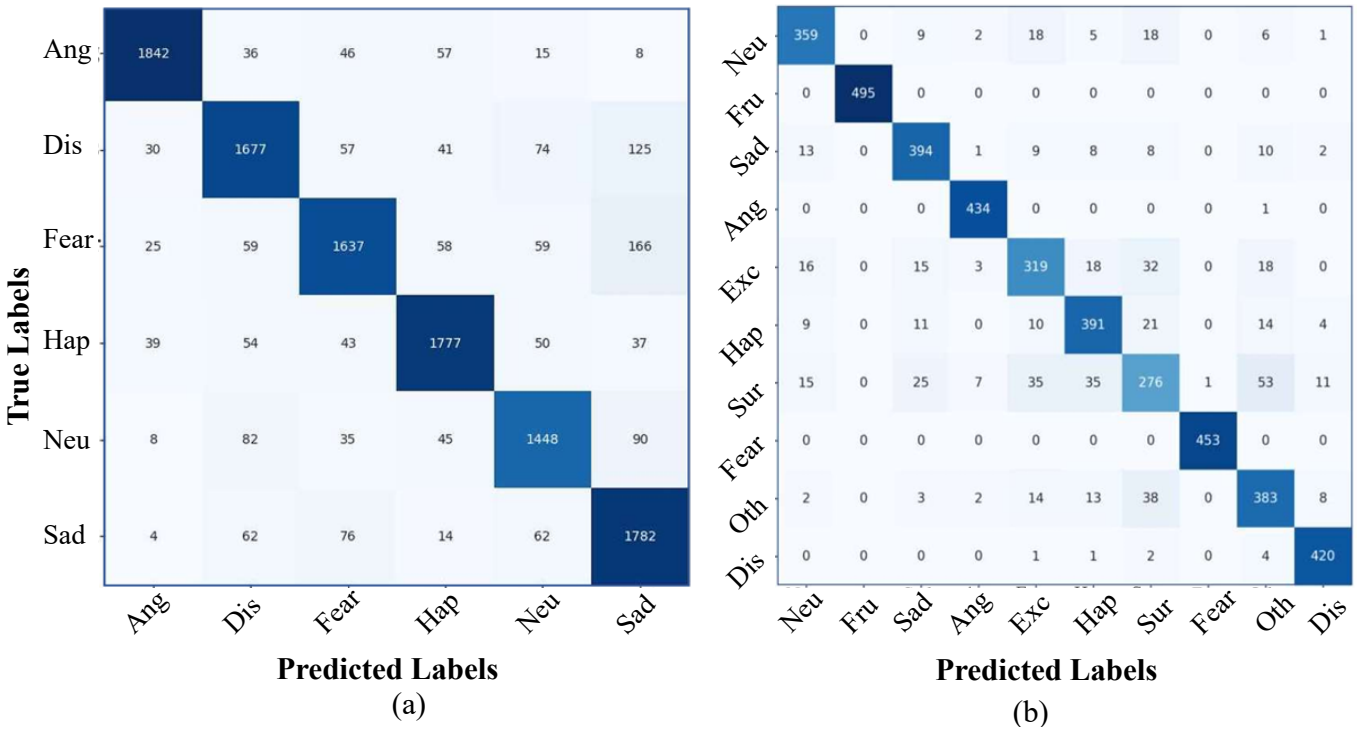


Fig. 3: Confusion matrices for emotion classification: (a) CREMA-D dataset, (b) IEMOCAP dataset.

overlap, insufficient data diversity, and limitations in capturing temporal emotional dynamics.

Table I also summarizes the performance of the baseline model and the proposed network on the IEMOCAP dataset, which was reduced to 7,460 samples covering all 10 emotions. The baseline model achieved an accuracy of 70%, while incorporating dilated convolutions increased the performance slightly to 84%, demonstrating the benefit of dilated CNNs in enhancing the model's ability to capture complex patterns in speech data.

When data augmentation was applied, the results showed notable improvement. The baseline EmoTech model achieved an accuracy of 71%, while the proposed model outperformed it with an accuracy of 87% (Table I). This underscores the importance of data augmentation in enhancing model generalization and classification performance.

The confusion matrix in Figure 3 provides a detailed breakdown of the model's performance across all emotion classes, highlighting its strengths . Here misclassifications occurred such as fear and surprise (53 cases), and happiness and excitement (32 cases), likely because of overlapping acoustic features such as pitch and intensity. Comparable emotional expressions, for example, rapid prosodic shifts in fear and surprise, are hard to distinguish.

The comparison between proposed and others baseline model is shown in table II.

The comparison reveals that the Proposed Model achieves the highest accuracy, with 86.00% on CREMA-D and 87.00% on IEMOCAP, surpassing all other models. EmoTech achieves 83.00% on CREMA-D and 71.00% on IEMOCAP, showing competitive performance. In contrast, ResNet-50 yields 68.12 % on CREMA-D , while GRU has the lowest accuracy of 55.01% on CREMA-D.

TABLE II: Comparison of Speech-Only Accuracy on CREMA-D and IEMOCAP Datasets

Model	CREMA-D (%)	IEMOCAP(%)
ResNet-50 [8]	68.12	–
GRU [12]	55.01	–
PATHOSnet v2 [7]	–	80.40
CAT [11]	–	73.80
EmoTech [3]	83.00	71.00
Proposed	**86.00**	**87.00**

B. Ablation Study

Table III presents the classification metrics—Precision, Recall, and F1-Score—for various models trained and evaluated on the IEMOCAP and CREMA-D datasets. The table enables a direct comparison of the performance of Support Vector Machine (SVM), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), and the proposed model.

The results reveal that overall model performance is better on the IEMOCAP dataset than on the CREMA-D dataset. The BiLSTM model achieves the highest F1-Score on both datasets, with a significant improvement on the IEMOCAP dataset (0.51) compared to CREMA-D (0.41). Among the tested models, the CNN model shows the weakest performance, particularly in terms of recall. In contrast, the pro-

posed model, incorporating dilated convolutions, demonstrates notable improvements in precision, recall, and F1-Score over the baseline, excelling at capturing complex temporal patterns in emotional speech. This analysis underscores the strengths and limitations of each approach while highlighting the adaptability and effectiveness of the proposed model across diverse datasets.

TABLE III: Ablation study of the proposed network using IEMOCAP and CREMA-D Datasets

Model	**IEMOCAP**			**CREMA-D**		
	P	**R**	**F1**	**P**	**R**	**F1**
SVM	0.67	0.36	0.47	0.41	0.41	0.40
CNN	0.61	0.23	0.34	0.41	0.35	0.34
LSTM	0.59	0.35	0.44	0.40	0.43	0.41
BiLSTM	0.64	0.42	0.51	0.41	0.42	0.41
EmoTech [3]	0.71	0.71	0.71	0.82	0.82	0.82
Proposed	**0.87**	**0.87**	**0.87**	**0.86**	**0.86**	**0.86**

IV. Conclusion

This study proposes a machine-learning algorithm for SER. The model considers the performance of various models and architectures in recognizing emotion in speech on the CREMA-D and IEMOCAP datasets. Tests indicate that the application of state-of-the-art methods, such as data augmentation, increases significantly and improvement in the architecture with dilated convolution layers further improves the accuracy of the emotion classification. On the CREMA-D dataset, the baseline EmoTech model achieved an average validation accuracy of 83%. The proposed model improved the accuracy to 86% after applying data augmentation. When tested on 10 balanced emotions with 7,460 samples, the enhanced version outperformed the EmoTech baseline in achieving a validation accuracy of 87% compared to the baseline's 86%. Those results have demonstrated the importance of data augmentation and advanced architectural features in making an emotion recognition system better. Future work should apply the method to more diversified datasets and emotional categories, especially in real-world noisy data situations. Most assuredly, integrating multimodal data—facial expression, gestures, and physiological signals—can be obtained in combination with speech. Cross-lingual studies and cross-cultural studies expand the knowledge base to find increasingly fine discrimination between universal patterns and the language-specificity of the emotional repertoire. Moreover, XAI approaches will lift model interpretability, providing clear assurance toward receiving trusting applications. These systems can be used in mental health, human-computer interaction, and affective computing.

References

[1] A. Badshah, J. Ahmad, N. Rahim, and S. W. Baik. Speech emotion recognition from spectrograms with deep convolutional neural network. In *2017 International Conference on Platform Technology and Service (PlatCon)*, pages 1–5. IEEE, 2017.

[2] Y. Bengio, A. Courville, and P. Vincent. Learning deep architectures for ai. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2017.

[3] S. Bin Habib Avro, T. Taher, and N. Mamun. Emotech: A multi-modal speech emotion recognition using multi-source low-level information with hybrid recurrent network. In *International Conference on Signal Processing, Information, Communication, and Systems*, pages 1–5. IEEE, 2024.

[4] L. Deng, Y. Zhao, and S. Renals. Speech emotion recognition with deep learning techniques. In *2019 IEEE Spoken Language Technology Workshop (SLT)*, pages 213–218, 2019.

[5] M. El Ayadi, M. S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[6] Rashedul Hasan, Meher Nigar, Nursadul Mamun, and Sayan Paul. Emotech: A multi-modal speech emotion recognition using multi-source low-level information with hybrid recurrent network. In *27th International Conference on Computer and Information Technology*, pages 1–5. IEEE, 2024.

[7] Y. He, N. Minematsu, and D. Saito. Multiple acoustic features speech emotion recognition using cross-attention transformer. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[8] N. C. Ristea and R. T. Ionescu. Self-paced ensemble learning for speech and audio classification. *arXiv preprint arXiv:2103.11988*, 2021.

[9] M. R. Sarker, M. M. Islam, and R. Hasan. Text-independent speech emotion recognition using hybrid deep neural networks. *IEEE Access*, 11:4567–4578, 2023.

[10] A. Satt, S. Rozenberg, and R. Hoory. Efficient emotion recognition from speech using deep learning on spectrograms. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2746–2750. IEEE, 2017.

[11] V. Scotti, F. Galati, L. Sbattella, and R. Tedesco. Combining deep and unsupervised features for multilingual speech emotion recognition. In *Proc. of the International Conference on Machine Learning and Data Mining (MLDM)*, 2021.

[12] A. Shukla, K. Vougioukas, P. Ma, S. Petridis, and M. Pantic. Visually guided self-supervised learning of speech representations. In *ICASSP 2020*, 2020.

[13] S. Tripathi and H. Beigi. Multi-task learning for emotion recognition from speech. In *Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 159–165, 2018.

[14] W. Wang, X. Li, and Y. Zhao. Dual-sequence lstm for speech emotion recognition with improved feature extraction. In *Proceedings of the 2020 International Conference on Artificial Intelligence and Robotics (ICAIR)*, pages 301–305, 2020.

[15] P. Yenigalla, P. Kumar, S. Pydi, and V. Aggarwal. Speech emotion recognition using spectrogram and phoneme embedding. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5099–5103, 2018.

[16] S. H. Yoon, S. Y. Byun, A. K. Dey, and J. H. Im. Speech emotion recognition using deep sparse autoencoder-based deep neural networks. *IEEE Transactions on Cognitive and Developmental Systems*, 10(3):557–568, 2018.