Speech Emotion Recognition from Audio Data Using LSTM Model

Md. Mahbub-Or-Rashid[1], Akash Kumar Nondi[2], Abdullah Al Sadnun[3],
Md. Anwar Hussen Wadud[4], T M Amir Ul Haque Bhuiyan[5], Md.Saddam Hossain[6]
Department of CSE, Bangladesh University of Business and Technology, Dhaka, Bangladesh[1, 2, 3, 4, 5, 6]
Department of CSE, Sunamgonj Science and Technology University (SSTU), Sunamganj, Bangladesh[4]

***Abstract*—The capacity to comprehend and interact with others through language is the most valuable human ability. Since emotions are crucial to communication, we are well-trained to recognize and interpret the many emotions we encounter. Contrary to popular assumption, the subjective aspect of human mood makes emotion recognition difficult for computers. There are some works based on Emotion recognition using images, text, and audio. We are here working on the audio dataset to find the accurate human emotion for computers to understand. In this work, we have utilized a Long Short-Term Memory (LSTM) model to implement Speech Emotion Recognition (SER) from Audio data on two different datasets: the Toronto Emotional Speech Set (TESS) and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). The accuracy rates of our LSTM-based model were impressive, with 91.25% for the RAVDESS dataset and 98.05% for the TESS dataset; the combined accuracy for both datasets was 87.66%. These results highlight the effectiveness of the LSTM model in effectively identifying and categorizing emotional states from audio files. The study adds significant knowledge to the field of speech emotion recognition by emphasizing the model's ability to handle a variety of datasets and its potential.**

Keywords*—Emotion; audio data; Ryerson audio-visual database; Toronto emotional speech set; classification; layers; combine

I. Introduction

Multimedia pattern recognition, especially Speech Emotion Recognition (SER), faces challenges due to emotional complexity. Deep learning simplifies the process by automatically learning features, unlike traditional methods requiring manual optimization [1]. Emotion recognition is essential for human-machine interaction but challenging due to fuzzy boundaries, individual expression, and overlapping emotions. Popular models like the circumflex of affect use dimensions like arousal and valence. Deep Neural Networks (DNNs), including CNNs and LSTMs, have advanced emotion recognition by leveraging features like Mel-Frequency Cepstral Coefficients (MFCCs). Early studies used handcrafted features, but this study proposes an end-to-end convolution-recurrent neural network that outperforms previous models on the RECOLA database [2]. This study explores speech emotion recognition (SER) using deep learning, focusing on Bangla and English datasets. The SUBESCO (Bangla) and RAVDESS (English) corpora were used, with mel-spectrograms as input features. A novel architecture, DCTFB (Deep CNN with Time-distributed Flatten and BLSTM layers), was proposed, achieving 86.86% accuracy on SUBESCO and outperforming existing models. The architecture combines local and sequential feature extraction for improved emotion detection. Additionally, a cross-lingual analysis using transfer learning and multi-corpus training was conducted, highlighting its effectiveness for SER across languages [3]. This study focuses on enhancing speech emotion recognition (SER) by addressing cost complexity and performance limitations. A lightweight CNN model with rectangular kernels and a modified pooling strategy is proposed to efficiently analyze frequency features in speech spectrograms. The model reduces parameters while maintaining high recognition accuracy, as demonstrated on IEMOCAP and EMO-DB datasets. Key contributions include improved computational efficiency and robust performance compared to baseline models, advancing SER for real-time applications in human-computer interaction [4].

Technology continues to evolve, significantly impacting human life and society. In human-machine interaction, speech is a key medium for communication, but recognizing emotions accurately remains a challenge for machines. While humans can easily detect emotions, machines struggle to interpret emotional states from speech. Speech, being the most natural form of communication, offers a fast and efficient method for interaction. Despite advancements in speech recognition, achieving natural interplay between humans and machines is still a work in progress. Speech emotion recognition (SER) holds potential for applications like customer service, education, forensics, medical analysis, and aiding individuals with disabilities. Integrating emotional recognition into voice assistants like Alexa, Siri, Google Assistant, and Cortana could make them more user-friendly and intuitive [5]. Accurately identifying emotional states in speech enhances human-machine interaction by enabling machines to respond appropriately. Speech emotion recognition (SER) has applications in areas like translation, education, customer service, and assistive technologies. Emotions can be classified into primary categories (e.g., Joy, Anger) or understood through cognitive perspectives. SER systems focus on extracting emotional cues from speech, often combined with facial expressions or biological features, to improve recognition accuracy. This study reviews SER features, systems, datasets, and experimental results, highlighting advancements in emotional speech recognition [6]. Deep learning enhances Speech Emotion Recognition (SER) by using CNNs for features and LSTMs for sequences. Turn-based SER often outperforms frame-based methods. This

study achieves state-of-the-art results on the IEMOCAP database, exploring deep learning models for simpler, more robust SER systems [7].

Speech Emotion Recognition (SER) detects emotions in speech for applications like human-computer interaction and robotics. It involves emotion modeling, preprocessing, feature extraction, and classification using deep learning. The study reviews these methods and current challenges in SER [8]. Speech Emotion Recognition (SER) helps improve human-computer interaction by recognizing emotions in speech. It has applications in smart speakers, virtual assistants, online education, therapy, safety, and customer service. Deep learning has significantly improved SER accuracy, surpassing traditional methods like HMMs, GMMs, and SVMs. SER training datasets can be natural, semi-natural, or simulated. Recent advancements include the use of recurrent neural networks (RNNs), LSTM, and auto-encoders. The study reviews existing datasets, traditional and deep learning methods, and suggests future directions for SER research [9]. Automatic Speech Emotion Recognition (SER) enables computers to understand human emotions through speech, enhancing natural communication. It uses methods like SVM and NN for classification, with the Berlin Emotional Database for training. SER is applied in areas like psychiatric diagnosis, educational software, and call center monitoring [10].

In [11], the authors focus on improving human-machine interaction by developing methods to identify emotions in speech. While emotions are often conveyed through voice, gestures, and body language, the study explores using Mel-recurrence Cepstral Coefficients to analyze sound for emotion recognition. Emotion recognition has broad applications in fields like brain research, psychiatry, and neuroscience, as well as in practical uses such as smart devices, customer service, and self-driving cars. The study discusses three types of training datasets for Speech Emotion Recognition: natural, semi-natural, and simulated. The key contributions outlined in the study are as follows:

- The CNN model achieved a 97.1% accuracy rate in categorizing emotional content of audio files.
- The deep learning models, particularly CNN, outperform the machine learning models in terms of accuracy, with an overall accuracy of 0.97. The sad class had the most outstanding performance, with a score of 0.90, and the worst on the surprised and disgust class (0.77).
- Using the RAVDESS and TESS datasets, proposed methods based on deep neural networks and machine learning to categorize emotions.
- Achieved an accuracy of 0.92, 0.93, and 0.971, respectively, using the deep learning model long short-term memories (LSTMs), GRU, and CNN.

Speech Emotion Recognition (SER) focuses on decoding emotions in spoken language using machine learning and AI. It has applications in areas like human-computer interaction and mental health diagnostics. The field explores how technology can understand and interpret emotional nuances in speech, offering insights into both technical methods and real-world impacts.

The remainder of this study is structured as follows**:** Section II provides the background and a detailed review of related works in the field of Speech Emotion Recognition (SER), focusing on both conventional and deep learning approaches. Section III outlines the methodology adopted in this research, including model architecture, data preprocessing techniques, and feature extraction processes. It discusses the datasets used in this study, specifically the RAVDESS and TESS datasets, along with data distribution and preprocessing details. It also explains the implementation details of the proposed LSTM model. Section IV presents the results, performance evaluations, and comparative analysis with existing methods. Finally, Section V concludes the study by highlighting the findings and proposing directions for future research.

II. Background Study and Related Works

A. Background

Using LSTM's prediction model along with machine learning and deep learning techniques to build the network training structure as well as make predictions. This underscores the objective of exploiting deep learning techniques developed through performance monitoring systems to compensate for incorrect predictions during the prediction process. There are a number of relevant sources cited in this study. The first source reviews the current literature relative to various theories and models. The second source emphasizes the realism of the research and illustrates various situations, including case studies and real events. In Source 3, the authors compare the various strategies previously employed with an assessment of their advantages and disadvantages. Various studies combine to form a strong foundation for this research. They reveal what is currently known about the topic and highlight the issues that will be addressed in this research [12].

B. Literature Survey

Limited progress in Bengali emotion recognition for sentiment analysis and abusive text detection. Methods include Chinese quarantine emotionally charged hotel reviews, and using machine learning for Covid-related content in Bangladesh. An example of such efforts is a Bengali article-based model for emotion detection as well as an advanced Bengali text annotation dataset for advanced emotion recognition [13]. In a cross-linguistic study involving Chinese and German, it was demonstrated that pitch, speech power and MFCC parameters are crucial for speech emotion detection. Alternatively, feature selection methods such as submodular functions and sequential forward selection are used with the Open Smile toolkit to extract a new feature set. The use of T-SNE revealed enhanced performance mainly in combining spectrogram and acoustic properties. Interestingly, this work specifically focuses on emotive character optimization for Bengali, thereby distinguishing it from previous studies [14]. Nevertheless, there are several issues related to speech emotion recognition, including the development of modern and advanced models leading to improved recognition that surpasses the state-of-the-art, as shown below. The modified

pooling technique combined with rectangular filters produces satisfactory performance with higher results than IEMOCAP (77.01%) and EMODB (92.02%). Frequency features are important SERs and suggested research directions for testing and deep learning in different databases [15]. Speech emotion recognition is reviewed in terms of different classifier models such as KNN, HMM, SSV, ANN and GMM. It uses various parameters like power, pitch, LPCC, etc., to recognize emotions. However, in capturing boundary nonlinearity, ANN, especially feedforward networks, show better results than GMM with the highest accuracy of 78.77%. HMM excels in temporal modeling, outperforms other classifiers, and SVM achieves perfect classification in multidimensional feature domains using different types of kernels [16]. A systematic review of Speech Emotion Recognition (SER) based on deep and traditional learning methods using available data sets. Some previous studies, such as Swain et al., 2018 and Khalil et al., 2019, were limited to conventional techniques or independent deep learning techniques, providing a shallow insight. However, Akcay et al. 2020 have been thoroughly researched on database, feature, classification and sentiment models applied to SER with emphasis on machine learning techniques [17]. Recognizing emotion from speech using basic technology for speech recognition. This study builds on previous studies showing that statistics of speech components (pitch, power, intonation, spectral shape, etc.) are associated with specific emotions. The results of the subjective assessment conducted using Interface Emotional Speech Synthesis are impressive - more than 80%. The work focused on low-level characterization and system design that could capture over eighty per cent recognition rates for seven emotions. Future studies are described, such as testing speaker/language-independent conditions and multi-modal emotion recognition [18]. To classify speech emotions (sadness, anger, fear and happiness), the study uses support vector machines after feeding features such as energy; MFCC coefficients (0 to 12) of Cocke-Younger algorithm output spectrum based on linear predictive coding model with pitch window size 35. Two classification methods, one-versus-all (OAA) and sex-based classification, are compared for females-only LPCC algorithm. As MFCC, time domain speech signal and various feature waveforms are extracted by research feature analysis. What's more, there are two datasets in the setting for testing - UGA (University of Georgia) and LDC (Linguistic Data Consortium). Among them are segmented recordings as well as student speech samples [19]. A detailed literature review on the effect of model size and pre-training data on downstream performance in Speech Emotion Recognition (SER) of the Transformer architecture. It compares different pre-trained variants of the wav2vec 2.0 and HuBERT models on arousal, dominance and valence dimensions across the MSP-Podcast dataset as well as the IEMOCAP and MOSI datasets. The main findings of the study include: that transformer models perform state-of-the-art in IEMOCAP; investigating why they are successful in improving valence spacing (excess avoidance), robustness and fairness, and efficiency. Authors publish their most successful models to allow the community to reproduce them [20]. By comparing Wav2Vec 2.0 for SER with V-FT and TAPT as baselines, we show that the latter outperforms TAPT in both accuracies on the IEMOCAP dataset using a novel approach known as P-TAPT. As it excels in low-resource settings and overall performance. The experimental framework is to evaluate SER on IEMOCAP and SAVEE datasets, taking frame-level emotion information into account and performing k-means clustering for better results. Research has pointed out that SER plays an important role in human-machine interaction and communication systems. It also shows how self-supervised pre-trained models can compensate for the lack of data annotation often associated with deep learning-based systems [21]. Furthermore, recent contributions have focused on leveraging machine learning and deep learning not only in emotion recognition but also in adjacent biomedical domains. For instance, Nurjahan et al. [22] present a comprehensive review of machine learning and deep learning algorithms in detecting COVID-19 through medical imaging, highlighting transferable modeling strategies and feature selection methods relevant to emotion recognition systems. Similarly, Bhuiyan et al. [23] developed a UNet-based deep learning segmentation framework for brain tumor detection, demonstrating the effectiveness of customized architectures in extracting domain-specific patterns — a methodology that may also enrich future SER frameworks tailored for low-resource languages like Bengali. In recent years, various approaches have been proposed to enhance emotion recognition from speech using deep learning and signal processing techniques. To enhance emotion recognition from speech using deep learning and signal processing techniques for speech emotion recognition (SER), demonstrating impressive accuracy rates on various datasets to find more accurate results using machine learning and deep learning algorithms, but their results are still not satisfactory. In another domain, the authors developed PCA-SIFT and weighted decision tree-based methods for diabetic retinopathy detection, emphasizing the importance of hybrid feature extraction and classification approaches for medical imaging applications [24]. These techniques can be conceptually adapted for SER feature extraction and classification. Additionally, their work on automatic water pump control using IoT [25] and an Android-based travel guide application [26] shows their practical expertise in building responsive systems—skills transferable to real-world emotion-aware systems. In the field of cloud computing, Mahbub-Or-Rashid et al. proposed resource-efficient frameworks for energy optimization and load balancing in cloud data centers [27][28], which are relevant when deploying computationally intensive SER models in scalable environments. The study on radio signal fading models [29] highlights the team's proficiency in signal analysis, a critical component in audio-based emotion recognition. Further contributions include a UNet-based deep learning framework for brain tumor segmentation [30] and a comprehensive review of ML/DL approaches for COVID-19 detection [31], both of which underscore expertise in medical image processing and transferable deep neural architectures. The development of Grainbee, a quantum-resistant blockchain system [32], and the FruVeg_MultiNet IoT system [33] reflect capabilities in secure, intelligent systems, suggesting possibilities for secure, privacy-preserving SER systems in the future. Moreover, the application of multi-output regression to predict death counts based on short-term mortality data [34] demonstrates

competence in statistical modeling, which could be integrated into emotion progression tracking in long-form audio content.

III. METHODOLOGY

The research presents a Speech Emotion Recognition (SER) model developed using Long Short-Term Memory (LSTM) networks. Leveraging LSTM's sequential learning capabilities, the model effectively identifies complex emotional patterns in audio data. It is designed for adaptability and generalization across datasets, specifically utilizing RAVDESS and TESS. The study emphasizes the model's contribution to SER research while avoiding detailed training metrics. Fig. 1 illustrates the data preprocessing and model development workflow.

A. Dataset

We use two well-known datasets, the Toronto Emotional Speech Set and RAVDESS from Kaggle, to perform a thorough analysis and pre-processing of the data. The Toronto Emotional Speech Set is a collection of emotionally charged speech recordings, and RAVDESS is a stable dataset containing a variety of emotional expressions. In order to capture relevant acoustic characteristics associated with emotional speech, the proposed model uses a thorough approach to data analysis and preprocessing in the domain of Speech Emotion Recognition (SER) from audio data using Long Short-Term Memory (LSTM) networks. A key component of this strategy is the use of Mel Frequency Cepstrum Coefficient (MFCC) as a feature extraction method. A comprehensive approach to data analysis and pre-processing highlights the model's ability to detect subtle nuances in emotional speech patterns, adding to the overall efficacy of the LSTM-based Speech Emotion Recognition system. The pre-processing phase carefully evaluates two diverse datasets, RAVDESS and TESS, to guarantee the model's flexibility and generalization across a range of emotional expressions. The Mel Frequency Cepstral Coefficients function as an essential set of features, offering a compact representation of the spectral characteristics in the audio signals. We have here 1200 audio data for RAVDESS and 1800 data for the TESS dataset. And the combined data is 3000. In Fig. 2 below, we can see the class distribution for all dataset.

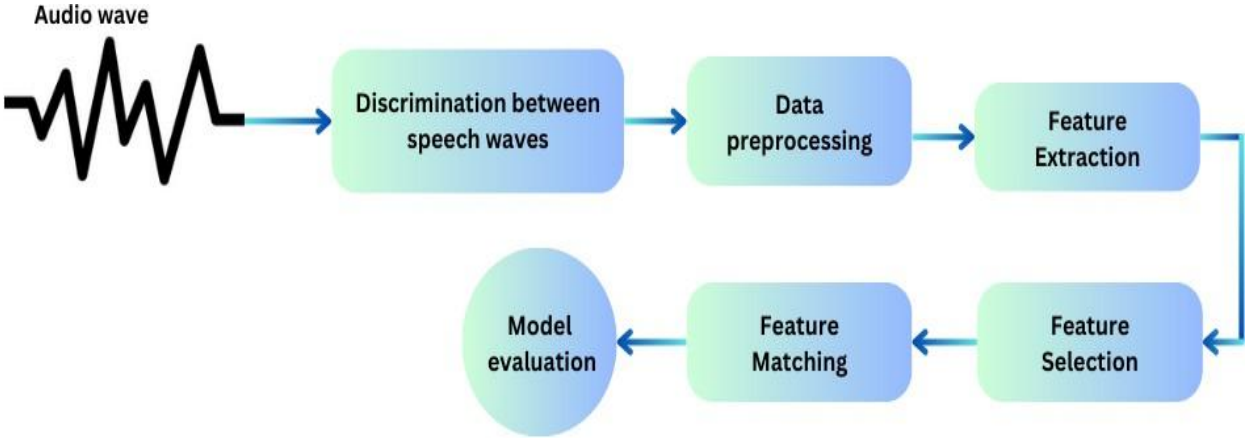
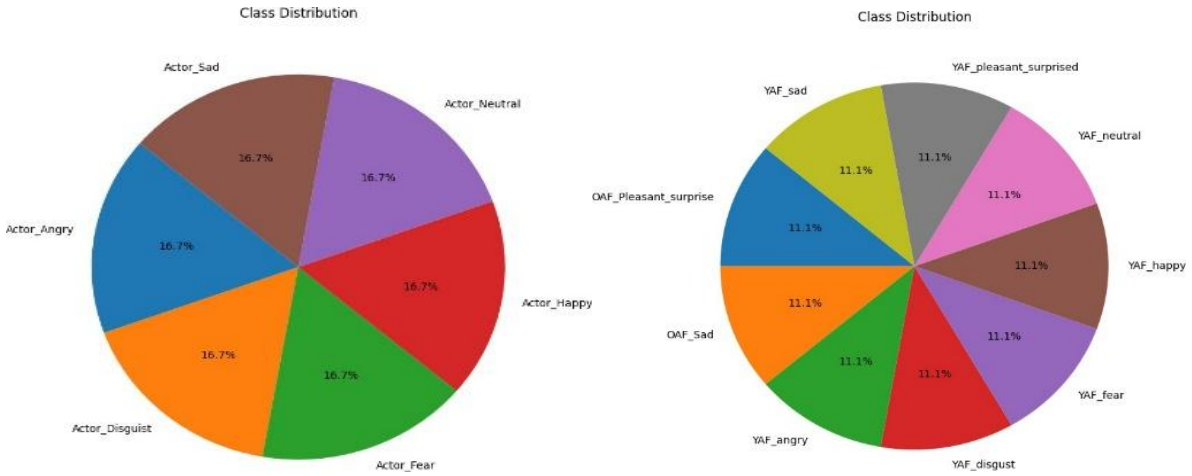


Fig. 1. Data preprocessing and model development workflow.



(a) RAVDESS. (b) TESS.

Fig. 2. Distribution of the dataset.

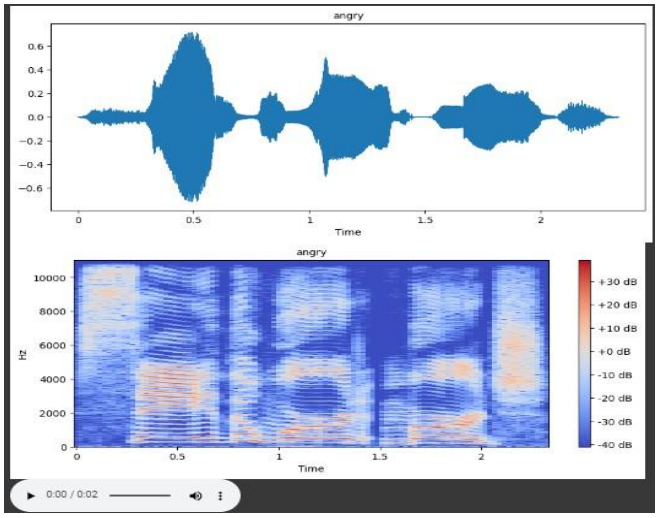


Fig. 3. Sample dataset of sad emotion.

B. Data Pre-processing

First, look at the audio signals as waveforms, which show how the sound changes over time. The "sad" waveform shows low energy and smooth changes (see Fig. 3), while the "angry" waveform has higher energy and sharp changes, showing strong emotions (see Fig. 4). Next, we transform the audio signals into spectrograms, which show the different frequencies in the sound over time. To improve the quality of the data, we may apply processes like noise reduction and normalization. The 'sad' spectrogram shows energy in lower frequencies with softer levels, while the 'angry' spectrogram has energy spread across a wider range of frequencies and higher intensities. These improved spectrograms help us better to capture the important patterns in the sound needed to identify emotions.

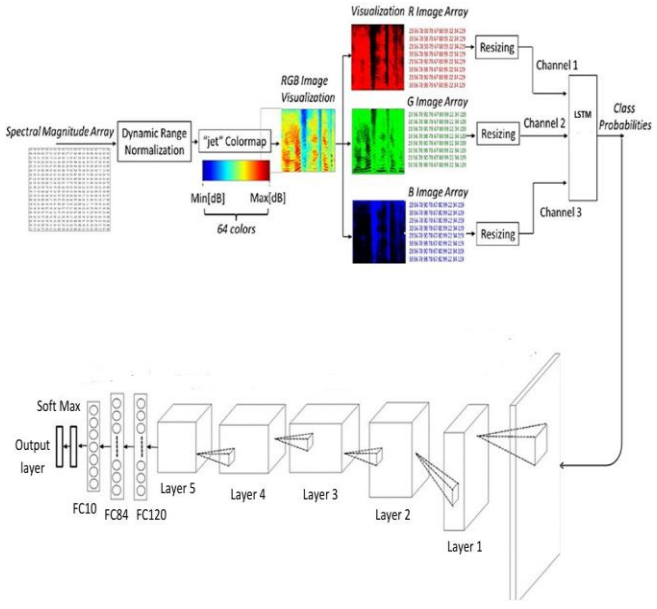


Fig. 4. Sample dataset of angry emotion.

Model implementation in our model for Speech Emotion Recognition begins by processing audio input into a spectral magnitude array, which represents the frequency content of the audio signal over time. To make the data consistent and enhance its usability, dynamic range normalization is applied. Next, the normalized spectrogram is transformed into an RGB image using the "jet" colormap, which assigns color values to the spectrogram, highlighting its features visually. The resulting RGB image is then split into three separate channels (Red, Green, and Blue), which are resized to a standard shape for further processing. These resized channels serve as input to an LSTM (Long Short- Term Memory) network (see Fig. 5). LSTM layers are specifically designed to capture sequential patterns in data, making them ideal for analyzing the temporal dynamics present in speech signals. After the LSTM processes the temporal information, the extracted features are passed through three fully connected (dense) layers (FC120, FC84, and FC10). These layers refine and condense the learned features to prepare them for classification. Finally, the output layer uses a softmax activation function to produce probabilities for each emotion class (e.g., happiness, sadness, anger), allowing the model to classify the speaker's emotional state based on their speech. This approach effectively combines audio preprocessing, visual representation, and deep learning to analyze and classify emotions in speech data. Comparative studies clarify the differences in performance between conventional machine learning algorithms and the LSTM model. This thorough investigation seeks to clarify the subtle benefits of utilizing deep learning architectures in SER, offering important information for the development of emotion-aware technology and (Human-Computer Interaction) HCI systems.

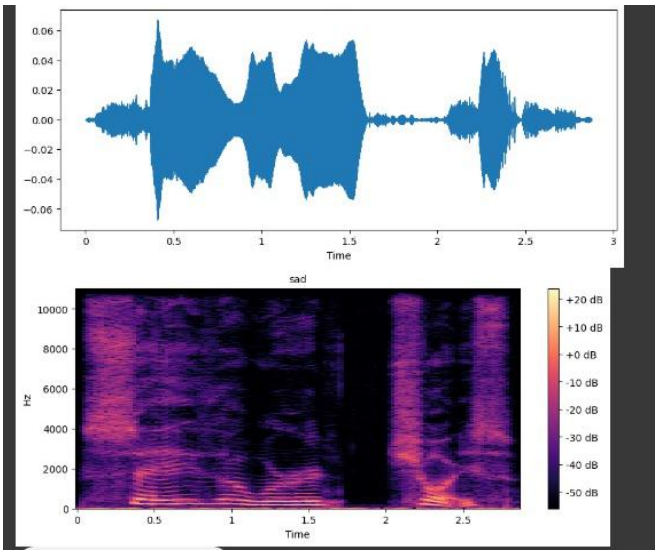


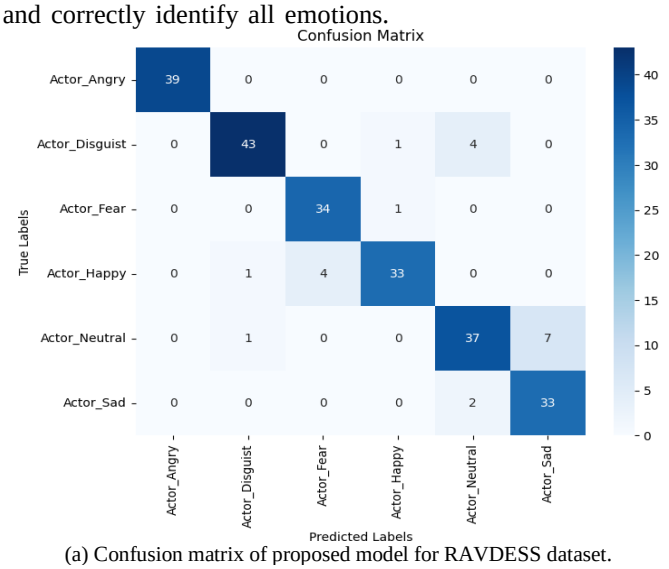
Fig. 5. Proposed Long Short-Term Memory (LSTM) model architecture.

IV. Result Analysis and Discussion

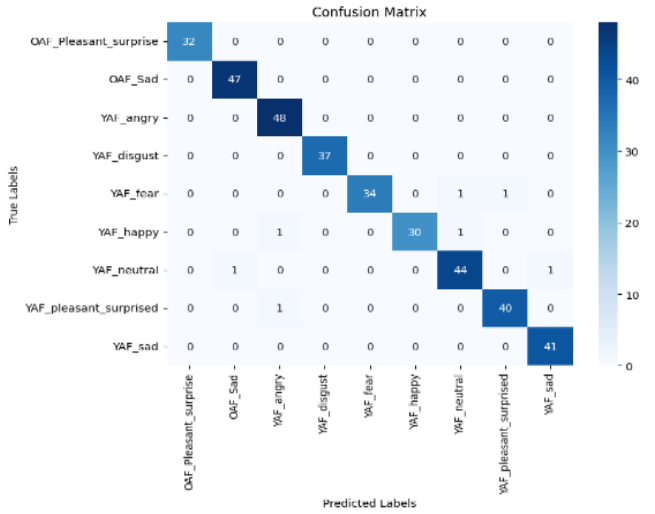
A. Result Analysis

1) Evaluation metrics: An illustration of a categorization process called a confusion matrix illustrates how closely the model's predictions match the actual effects (see Fig. 6). We used a total of 240 voice data for RAVDESS, 360 voice data for TESS and 600 voice data for combined RAVDESS and TESS to test the proposed model LSTM. Fig. 6(a), here 240

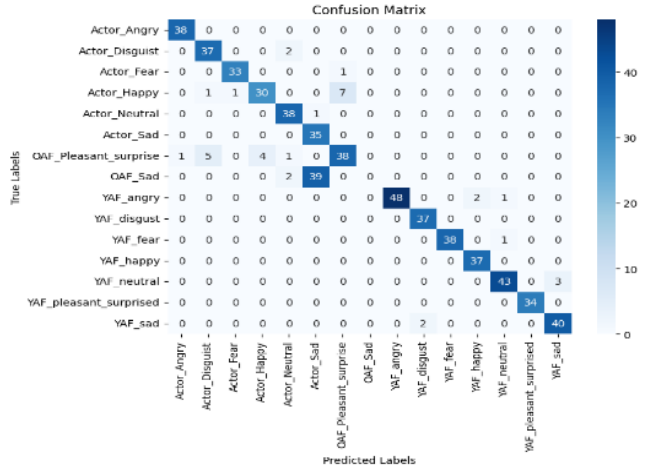
voice data from RAVDESS, Fig. 6(b), here 360 voice data from TESS and Fig. 6(c), here 600 voice data from combined RAVDESS and TESS to test the proposed model LSTM and correctly identify all emotions.



(a) Confusion matrix of proposed model for RAVDESS dataset.



(b) Confusion matrix of proposed model for TESS dataset.



(c) Confusion matrix of proposed model for combined RAVDESS and TESS dataset.

Fig. 6. Confusion matrix of proposed model.

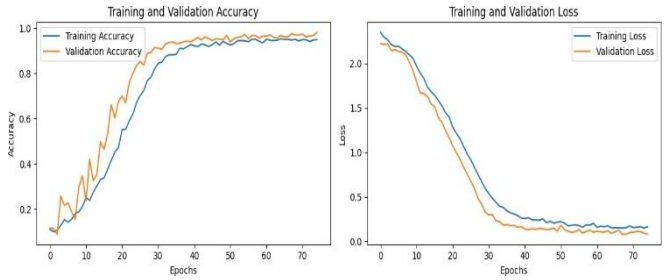


Fig. 7. Accuracy of the proposed LSTM model for RAVDESS dataset.

B. *Performance Evaluation*

The calculation was done using the Keras callbacks method. While experimenting with various epoch counts, we measured the precision for both training and validation. The model reaches its peak accuracy in training, testing, and verification after 75 epochs (see Fig. 7 to Fig. 9). In the graphs, these results could be attributed to the rich and diverse nature of the TESS and RAVDESS datasets, which have contributed to the model's robust training. These outcomes highlight the effectiveness of the model in accurately classifying emotions within speech signals. The high accuracy rates indicate the successful capture of temporal dependencies.

Fig. 8. Accuracy of the proposed LSTM model for TESS dataset.

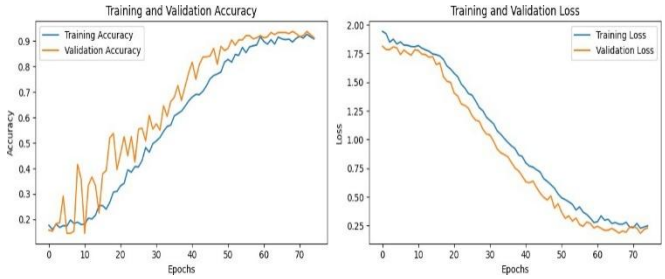


Fig. 9. Accuracy of the proposed LSTM model for combined RAVDESS and TESS dataset.

This study discusses the performance of the LSTM model that has been implemented and trained. Data collected from the RAVDESS and Toronto Emotional Speech Set (TESS) dataset, which includes 1200 voice data of RAVDESS and 1800 voice data of TESS, were used to test the model presented in this research. Given the skewness of the dataset, we went beyond simple classification accuracy to assess model performance by using other metrics, such as precision, sensitivity, recall, and F1-score, which incorporates the metrics with Fig. 10.

Classification Report:

	precision	recall	f1-score	support
Actor_Angry	1.00	1.00	1.00	39
Actor_Disguist	0.96	0.90	0.92	48
Actor_Fear	0.89	0.97	0.93	35
Actor_Happy	0.94	0.87	0.90	38
Actor_Neutral	0.86	0.82	0.84	45
Actor_Sad	0.82	0.94	0.88	35
accuracy			0.91	240
macro avg	0.91	0.92	0.91	240
weighted avg	0.92	0.91	0.91	240

Number of training samples: 960
Number of testing samples: 240

(a) Classification report of proposed model for RAVDESS dataset.

Classification Report:

	precision	recall	f1-score	support
OAF_Pleasant_surprise	1.00	1.00	1.00	32
OAF_Sad	0.98	1.00	0.99	47
YAF_angry	0.96	1.00	0.98	48
YAF_disgust	1.00	1.00	1.00	37
YAF_fear	1.00	0.94	0.97	36
YAF_happy	1.00	0.94	0.97	32
YAF_neutral	0.96	0.96	0.96	46
YAF_pleasant_surprised	0.98	0.98	0.98	41
YAF_sad	0.98	1.00	0.99	41
accuracy			0.98	360
macro avg	0.98	0.98	0.98	360
weighted avg	0.98	0.98	0.98	360

Number of training samples: 1440
Number of testing samples: 360

(b) Classification report of proposed model for TESS dataset

Classification Report:

	precision	recall	f1-score	support
Actor_Angry	0.97	1.00	0.99	38
Actor_Disguist	0.86	0.95	0.90	39
Actor_Fear	0.97	0.97	0.97	34
Actor_Happy	0.88	0.77	0.82	39
Actor_Neutral	0.88	0.97	0.93	39
Actor_Sad	0.47	1.00	0.64	35
OAF_Pleasant_surprise	0.83	0.78	0.80	49
OAF_Sad	0.00	0.00	0.00	41
YAF_angry	1.00	0.94	0.97	51
YAF_disgust	0.95	1.00	0.97	37
YAF_fear	1.00	0.97	0.99	39
YAF_happy	0.95	1.00	0.97	37
YAF_neutral	0.96	0.93	0.95	46
YAF_pleasant_surprised	1.00	1.00	1.00	34
YAF_sad	0.93	0.95	0.94	42
accuracy			0.88	600
macro avg	0.84	0.88	0.86	600
weighted avg	0.84	0.88	0.85	600

Number of training samples: 2400
Number of testing samples: 600

(c) Classification report of proposed model for combined RAVDESS and TESS dataset

Fig. 10. Classification report of proposed model.

1) Comparison with existing works: The classification outcomes of conventional machine learning techniques are shown in Table I, alongside our suggested LSTM model. Our proposed model outperforms the other classifiers with 98.05% and 91.25% and a combine result of 87.66% accuracy compared to other ML classifiers mentioned.

TABLE I. PERFORMANCE COMPARISON WITH THE EXISTING METHODS

Methods	Algorithm	Dataset	Classification Accuracy
Issa et al. [11]	CNN	RAVDESS	71.61%
Zeng et al.[15]	CNN+LSTM	RAVDESS	64.48%
Zamil et al. [19]	GRU	RAVDESS	67.14%
Hashemzehi et al. [17]	SVM	RAVDESS	77.32%
Praseetha et al. [21]	CNN	TESS	95.82%
Ravi et al.[8]	CNN	TESS	97.1%
Huang et al. [19]	MLP	TESS	85%
Dupuis et al. [5]	LSTM	TESS	82%
Proposed	**LSTM**	RAVDESS	**91.25%**
Proposed	**LSTM**	TESS	**98.05%**
Proposed	**LSTM**	Combined RAVDESS and TESS	**87.66%**

This study demonstrates the impressive performance of the LSTM model in Speech Emotion Recognition (SER), achieving accuracy rates of 91.25% and 98.05% on the RAVDESS and TESS datasets, respectively, and 87.66% on the combined dataset. The model's ability to capture temporal dependencies and identify emotional nuances in speech signals highlights its effectiveness. Future research can focus on improving model generalization with larger, diverse datasets, integrating real-world scenarios, exploring ensemble or hybrid architectures, applying transfer learning, and enhancing robustness across languages, thus advancing SER methodologies.

V. CONCLUSION

In summary, we applied LSTM models for Speech Emotion Recognition (SER) using the RAVDESS and TESS datasets, achieving strong accuracy rates of 91.25%, 98.05%, and 87.66% for the individual and combined datasets. The results demonstrate the LSTM model's ability to effectively capture the emotional subtleties in speech and recognize temporal patterns within the data. By training on diverse and comprehensive datasets, the model was able to generalize well and deliver accurate emotion classifications. This research emphasizes the practical potential of LSTM models in real-world SER applications, such as human-computer interaction, customer service, and healthcare. Our findings provide a solid foundation for further development of emotion recognition systems that can be used across different domains, paving the way for more efficient and accurate emotion detection in speech.

REFERENCES

[1] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA)*. IEEE, 2016, pp. 1–4.

[2] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5089–5093.

[3] S. Sultana, M. Z. Iqbal, M. R. Selim, M. M. Rashid, and M. S. Rahman, "Bangla speech emotion recognition and cross-lingual study using deep cnn and blstm networks," *IEEE Access*, vol. 10, pp. 564–578, 2021.

[4] T. Anvarjon, Mustaqeem, and S. Kwon, "Deep-net: A lightweight cnn-based speech emotion recognition system using deep frequency features," *Sensors*, vol. 20, no. 18, p. 5212, 2020.

[5] J. Devnath, S. Hossain, M. Rahman, H. Saha, M. A. Habib, and N. Sultan, "Emotion recognition from isolated bengali speech," 2020.

[6] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden markov models," *Speech communication*, vol. 41, no. 4, pp. 603–623, 2003.

[7] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[8] M. B. Akc¸ay and K. Og˘uz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[9] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models," *Sensors*, vol. 21, no. 4, p. 1249, 2021.

[10] Y. Chavhan, M. Dhore, and P. Yesaware, "Speech emotion recognition using support vector machine," *International Journal of Computer Applications*, vol. 1, no. 20, pp. 6–9, 2010.

[11] R. R. Choudhary, G. Meena, and K. K. Mohbey, "Speech emotion based sentiment recognition using deep neural networks," in *Journal of Physics: Conference Series*, vol. 2236, no. 1. IOP Publishing, 2022, p. 012003.

[12] M. Ahmed, P. C. Shill, K. Islam, M. A. S. Mollah, and M. Akhand, "Acoustic modeling using deep belief network for bangla speech recognition," in *2015 18th international conference on computer and information technology (ICCIT)*. IEEE, 2015, pp. 306–311.

[13] T. Ahmed, S. F. Mukta, T. Al Mahmud, S. Al Hasan, and M. G. Hussain, "Bangla text emotion classification using lr, mnb and mlp with tf-idf & countvectorizer," in *2022 26th International Computer Science and Engineering Conference (ICSEC)*. IEEE, 2022, pp. 275–280.

[14] S. Sultana and M. S. Rahman, "Acoustic feature analysis and optimization for bangla speech emotion recognition," *Acoustical Science and Technology*, vol. 44, no. 3, pp. 157–166, 2023.

[15] R. D. G. Ayon, M. S. Rabbi, U. Habiba, and M. Hasana, "Bangla speech emotion detection using machine learning ensemble methods."

[16] A. B. Ingale and D. Chaudhari, "Speech emotion recognition," *International Journal of Soft Computing and Engineering (IJSCE)*, vol. 2, no. 1, pp. 235–238, 2012.

[17] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using cnn," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 801–804.

[18] A. Nogueiras, A. Moreno, A. Bonafonte, and J. B. Marin˜o, "Speech emotion recognition using hidden markov models," in *Seventh European conference on speech communication and technology*, 2001.

[19] M. Jain, S. Narayan, P. Balaji, A. Bhowmick, R. K. Muthu *et al.*, "Speech emotion recognition using support vector machine," *arXiv preprint arXiv:2002.07590*, 2020.

[20] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[21] L.-W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[22] Nurjahan, Md. Mahbub-Or-Rashid, Md. Shahriare Satu, Sanjana Ruhani Tammim, Farhana Akter Sunny, Mohammad Ali Moni "Machine learning and deep learning algorithms in detecting COVID-19 utilizing medical images: a comprehensive review" Iran Journal of Computer Science, 2024.Springer Nature.

[23] TM Amir-Ul-Haque Bhuiyan,Md Anwar Hussen Wadud, Md. Mahbub-Or-Rashid, Md. Reazul Islam "Brain Tumor Detection by Image Segmentation Using Customized UNet Deep Learning Based Model" 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT).

[24] F. T. Johora, M. Mahbub-Or-Rashid, M. A. Yousuf, T. R. Saha, and B. Ahmed, "Diabetic retinopathy detection using PCA-SIFT and weighted decision tree," in *Proc. Int. Joint Conf. on Computational Intelligence*, 2020.

[25] Md. Reazul Islam and M. Mahbub-Or-Rashid, "Design and Implementation of Cost Effective Automatic Water Pump Controlling System for Domestic Application using IoT," in *Proc. Conf. on Big Data, IoT and Machine Learning (BIM)*, 2023.

[26] J. Akther, M. Harun-Or-Roshid, M. Mahbub-Or-Rashid, and S. J. Soheli, "Bangladesh Travel Guide (BTG): An Android Mobile Application," *J. Adv. Res. Mobile Comput.*, vol. 3, no. 1, pp. 1–5, 2021.

[27] J. A. Jeba, S. Roy, M. O. Rashid, S. T. Atik, and M. Whaiduzzaman, "Towards green cloud computing: an algorithmic approach for energy minimization in cloud data centers," in *Research Anthology on Architectures, Frameworks, and Integration Strategies for Distributed and Cloud Computing*, IGI Global, pp. 1567–1589, 2018.

[28] M. Mahbub-Or-Rashid, M. J. N. Mahi, J. A. Jeba, and F. T. Johora, "Achieving an efficient approach through using resource allocation, management and load balancing for cloud data centers," *Recent Trends in Cloud Computing and Web Engineering*, vol. 3, no. 1, pp. 1–23, 2021.

[29] L. Nahar, M. Mahbub-Or-Rashid, S. Akter, and R. T. Khan, "Medium Access Probability of Cognitive Radio Network Under ECC-33/Hata-Okumura Extended Model Using Different Fading Channels at 1900MHz and 2100MHz," *Int. J. Comput. Eng. Res. (IJCER)*, vol. 6, no. 7, pp. 35–42, 2016.

[30] T. M. Amir-Ul-Haque Bhuiyan, M. A. H. Wadud, M. Mahbub-Or-Rashid, and M. Reazul Islam, "Brain Tumor Detection by Image Segmentation Using Customized UNet Deep Learning Based Model," in *Proc. 6th Int. Conf. on Electrical Engineering and Information & Communication Technology (ICEEICT)*, 2024.

[31] Nurjahan, M. Mahbub-Or-Rashid, M. S. Satu, S. R. Tammim, and F. A. Sunny, "Machine learning and deep learning algorithms in detecting COVID-19 utilizing medical images: a comprehensive review," *Iran J. Comput. Sci.*, vol. 4, no. 1, pp. 23–45, 2024.

[32] S. A. Joni, R. Rahat, N. Tasnin, P. Ghose, and M. Mahbub-Or-Rashid, "Grainbee: A Quantum-Resistant Blockchain-Based Ration Distribution System with Hardware Security Modules," in *Proc. IEEE Int. Conf. on Computing, Applications and Systems (ICCAS)*, 2024.

[33] T. M. A. U. H. Bhuiyan, M. Reazul Islam, J. I. Babar, M. A. Nur, and M. Mahbub-Or-Rashid, "FruVeg_MultiNet: A hybrid deep learning-enabled IoT system for fresh fruit and vegetable identification with web interface and customized blind glasses for visually impaired people," *J. Agric. Food Res.*, vol. 1, pp. 100–112, 2025.

[34] M. I. Ahmed, Nurjahan, M. Mahbub-Or-Rashid, and F. Islam, "Prediction of Death Counts Based on Short-term Mortality Fluctuations Data Series using Multi-output Regression Models," *Int. J. Adv. Comput. Sci. Appl. (IJACSA)*, vol. 14, no. 5, pp. 88–94, 2023.