iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.52422

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Attention-based Speech Emotion Recognition Approach for Medical Application

Pushkar Kane[1], Pratik Mathe[2], Yash Waghumbare[3], Ganesh Pise[4]
Department of Information Technology, Pune Institute of Computer Technology, Pune

Abstract: ***Presently AI is used in various medical fields. Mental health is an important part of the overall health of a person and speech is the primary form for expression of emotion. Thus, Speech Emotion Recognition can be used to understand emotions of the person and help doctors focus on the cure. Speech Emotion recognition is analysis and classification of speech signals to detect the underlying emotions. This paper proposes a model for speech emotion recognition using an attention mechanism. The model is developed using Mel-frequency cepstral coefficients (MFCCs), and a combination of 2D CNN layers and LSTM recurrent layers for temporal aggregation. The proposed model is evaluated using a dataset of speech recordings containing eight emotion categories. The results show that the model achieves 89% accuracy. The attention mechanism is found to improve the recognition performance by focusing on relevant emotional information and ignoring irrelevant information. This research has potential applications in clinical settings in detection as well as treatment for mental health issues.***

Keywords: Speech Emotion Recognition (SER), Mental Health, Deep Learning, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Mel Frequency Cepstrum Coefficients (MFCC), Attention Mechanism, RAVDESS

I. INTRODUCTION

Monitoring and treating mental health are crucial for overall physical health and a safer community and social life [1], as overall globally over 1.1 billon individuals were diagnosed with mental disorders in 2016 [2]. The COVID-19 pandemic has exacerbated the situation, with depression and anxiety disorders increasing by 25% globally during the first year, particularly among young people and women [3]. Unfortunately, the late or unreceived mental care has led to an increase in suicides, with one person dying by suicidal action related to a mental disorder every 40 seconds [4]. Furthermore, the United States has experienced over 200 cases of mass shootings in less than the first half of the year [5].

In speech, individuals express common emotions such as happiness, sadness, anger, worry, fear, and neutrality. Changes in an individual's emotions can be indicative of certain mental disorders such as depression, mood disorders and trauma and stress disorders. Early diagnosis of mental disorders is crucial to provide the correct treatment and prevent severe illnesses and suicidal actions. However, the current screenings for these diseases rely mainly on psychological examination and interview, which can lack objectivity. Therefore, there is a need for technology that can detect psychiatric changes in patients earlier. A therapist needs to understand the patients mental state by recognizing emotions through their vocal responses. Based on this, therapists then need to decide the course of treatment by analyzing the patients progress by comparing his responses in earlier therapeutic settings. Here it becomes challenging of the therapist to keep track of the progress of multiple patients and quantify the improvements in their responses. Thus, a platform can be useful to provide insights and keep track of patient emotional progress by recognizing the embedded emotions in patients' speech in various medical fields and therapy sessions.

AI can relieve doctors of the burden of comprehending their patients' emotional states and move the emphasis from transactional chores to individualized medical treatment and service. However, it necessitates that computers cleverly deduce human speech and comprehend it on a semantic level. Currently, artificial intelligence (AI) is empowering a variety of medical applications, including early detection of diseases, drug discovery, heart disease prediction and Robot-assisted surgery. Understanding emotions can improve human machine interaction by enabling machine to understand human behavior better and respond accordingly. Thus, various efforts are being taken in this domain.

SER is the quickest means of communication and information exchange between people and computers, and it has a various practical applications in human computer interaction. Systems for detecting the embedded emotions in voice signals are referred to as Speech Emotion Recognition (SER) systems. Speech Emotion Recognition uses extracted features from raw audio waves using various techniques like MFCC and Log-Mel-Spectrogram. These features can be temporal or spectral features. In this paper, we propose the use of MFCCs technique for feature extraction. The primary goal of SER systems is to identify certain speaker voice traits under various emotional states.

Speech emotion recognition technology can be used to detect, monitor and treat mental disorders such as depression, trauma, stress and bipolar disorder. By analyzing a patient's speech patterns, the technology can identify signs of emotional distress, which can be used to inform treatment and intervention plans. Speech emotion recognition technology can also be used to diagnose autism spectrum disorder (ASD). Individuals affected with autism often have difficulty recognizing and interpreting emotions in speech, and this technology can help clinicians assess these abilities in a standardized and objective way. Parkinson's disease can affect speech patterns, leading to changes in pitch, volume, and articulation. Speech emotion recognition technology can be applied to detect these changes, helping in early detection and treatment of Parkinson's disease. Individuals who have suffered a stroke may experience speech difficulties, including problems with emotional expression. Speech emotion recognition technology can also be employed in this case for monitoring progress during stroke rehabilitation and providing targeted interventions to improve emotional expression in speech. Overall, speech emotion recognition technology has the potential to revolutionize clinical diagnosis and treatment by providing objective, standardized assessments of emotional expression in speech.

The objective of this research is to develop a model for accurately detecting and classifying emotions from speech signals using deep learning techniques incorporating attention mechanisms in the model to improve recognition performance by focusing on relevant emotional information and ignoring irrelevant information. Further, we evaluate the built model and compare it with existing model's performances and explore the potential applications of the proposed model in clinical settings for the detection and treatment of mood disorders. We also focus on contributing to the development of AI-based tools for personalized care and medical services, thereby freeing up physicians' tasks of understanding the emotional space of their patients

II. BACKGROUND

One of the main challenges in SER is the variability of emotions conveyed through speech, which can be influenced by various factors such as cultural background, gender, age, and speaking style. In addition, there is often a lack of labeled speech data for emotion recognition, which makes it difficult to train accurate machine learning models. There are several approaches to feature extraction for SER, including spectral, prosodic, and lexical features. Spectral features involve analyzing the frequency spectrum of speech signals, while prosodic features involve analyzing the rhythm, intonation, and stress patterns of speech. Lexical features involve analyzing the content and language of speech.

Prior to the advent of deep learning, speech emotion recognition (SER) research relied heavily on the use of manually engineered audio features and old machine learning models such as Hidden Markov Models (HMMs) and support vector machines (SVMs) [6]. K. Han along with his colleagues introduced the first deep learning-based model [7] for SER in 2014. Their approach utilized deep neural networks to automatically extract complex features from the data, demonstrating the effectiveness of this method for SER. Specifically, they utilized Mel-frequency cepstral coefficients (MFCCs) to extract features and classified audio files into five different emotions. In [8] various statistical features as well as Mel frequency cepstral coefficients (MFCC) from a database of 2000 utterances are explored. The pitch feature was extracted using AMDF and employed a Naive Bayes classifier for classifying audio signals.

In [9] the author utilized recurrent neural networks (RNNs) on the IEMOCAP dataset and employed raw and emotional low-level descriptors (LLDs) for feature extraction. The resulting accuracy rate was found to be 66.23%, on average. Fatemeh Noroozi's research in 2017[10] presented a vocal-based approach that utilized decision trees and random forest algorithms. The average accuracy rate achieved by this method was 66.28%.

In 2017, a study on speech emotion recognition using spectrograms with deep convolutional neural networks (CNNs) was conducted.

The proposed CNN model consisted of three convolutional layers and three fully connected layers.[11]. The authors of [12] have investigated a model which integrates both temporal and spatial features. They expanded the RAVDESS dataset four-fold by adding Additive Gaussian White Noise (AWGN) to 5760 audio samples. To classify emotions from one of the eight classes, the authors created two parallel CNNs for extracting both spatial as well as temporal elements.

MFCC is commonly employed for analyzing audio signals and have shown superior performance for speech-centered emotion recognition methods in comparison to other techniques. In 2020[13], Mustaqeem introduced a CNN model with a deep bidirectional long short-term memory (LSTM) that incorporated MFCC and outperformed the existing models on the IEMOCAP dataset. To ascertain the performance of the advanced DSCNN and CNN models[14], the author conducted two experiments. The experiments resulted in an accuracy rate of 79.4%.

Speech emotion recognition technology can be used to diagnose mental issues of depression, stress, and bipolar disorder. In [15] authors have concluded that by analyzing a patient's speech patterns, the technology can identify signs of emotional distress, which can be used to inform treatment and intervention plans. Parkinson's disease can affect speech patterns, leading to changes in pitch, volume, and articulation. In [16] Tsanas have described that SER technology can be employed to detect these changes, providing early detection and treatment of Parkinson's disease. Authors in [17] demonstrate that individuals who have suffered a stroke may experience speech difficulties, including problems with emotional expression. SER can also be applied in monitoring progress during stroke rehabilitation and provide targeted interventions to improve emotional expression in speech. PTSD can affect speech patterns, leading to changes in tone, pitch, and volume. In [18] "Using interpretable machine learning models to improve PTSD diagnosis" authors suggest that SER technology can be utilized in monitoring improvement during PTSD treatment and provide targeted interventions to improve emotional expression in speech.

Overall, the field of SER is still evolving, and there is much room for more research and experimentation for improving accuracy and making emotion recognition systems more reliable. However, the potential applications of SER in various fields make it an area of research that is likely to continue to receive significant attention in the future. Speech emotion recognition technology has the potential to revolutionize clinical diagnosis and treatment by providing objective, standardized assessments of emotional expression in speech

III. PROPOSED METHODOLOGY

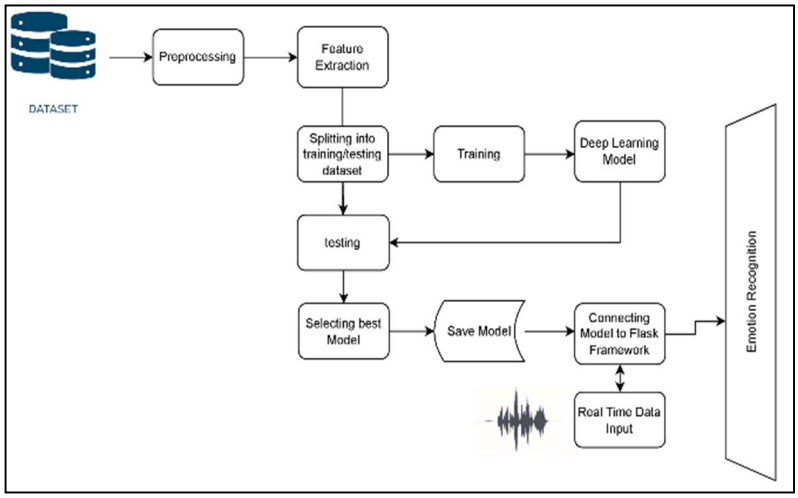


Fig.1 Overview of the proposed approach

A. Feature Extraction

To help the model learn to distinguish between audio files, features are extracted from the audio files. Figure 1 illustrates the overview of the suggested approach. Audio signals cannot be used as input to the model in raw format due to the presence of noise. Thus, feature extraction from the audio signals is essential. The most widely employed technique for feature extraction is MFCC [19]. The feature extraction process is performed using the Librosa library in Python, which is widely used for audio analysis. This library allows the visualization of audio signals and the use of different signal-processing techniques for feature extraction. The MFCC method is highly effective in providing a human-like perception of voice and achieving high accuracy in speech emotion recognition. MFCC reduces computational complexity, improves feature extraction, and enables the identification of parameters such as pitch and energy.

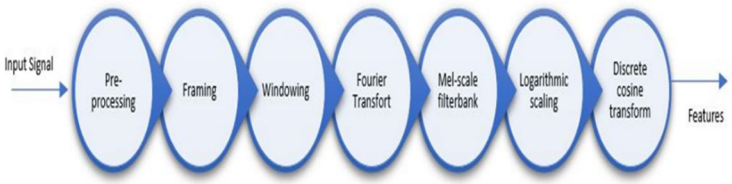


Fig. 2 MFCC generation

Fig. 2 depicts the diagram for the feature extraction process. During the A/D conversion process, the analog audio signal is transformed into a digital format with a sampling frequency of 16kHz. In order to process speech signals, they are divided into short time intervals known as frames, typically ranging from 20 to 40 ms. However, simply chopping the signal at the edges can result in noise owing to the sudden fall in amplitude. To avoid this, Hamming or Hanning windows are used instead of rectangular windows to reduce spectral distortion and minimize discontinuities at the edges of each frame.

The process of applying Discrete Fourier transform (DFT) using FFT, the audio signal is transformed into the frequency domain from the time domain. This transformation is beneficial because analyzing signals in the frequency domain is easier than in the time domain. The representation of a signal over time is depicted in a time domain graph, whereas a frequency domain graph illustrates the distribution of the signal across different frequency bands. Fast Fourier transform (FFT) is used to calculate the DFT of the digital audio signal sequence. To account for the fact that humans perceive audio frequency differently, Mel-filter bank is used to convert a given frequency to a frequency that corresponds to the way our ears perceive them. This is because humans are less sensitive to changes in audio signal energy at higher frequencies than at lower ones. The log function also exhibits similar properties: at lower input values, the gradient of the log function is higher, while at higher input values, the gradient value is lower. Therefore, we apply the log function to the output of the Mel filter to simulate the behavior of the human hearing system.

During the IDFT step, we perform the inverse transform for the Mel scale Filter bank output. The cepstrum, which is the inverse of the log of the magnitude of the signal, is used to obtain 12 features for the given signal, and sample energy is the 13th feature of MFCC.

In addition to the 13 features obtained using MFCC, the technique also considers their first and second-order derivatives, making more 26 features. Overall, the MFCC technique extracts a total of 39 features.

B. Classification Models

CNN: Convolutional neural networks (CNNs or ConvNets) are a type of neural network architecture which understands directly from raw data, without manual feature extraction. They are particularly effective at recognizing patterns in images, such as objects, faces, and scenes. As shown in Figure 3, a CNN contains an input layer, a convolutional layer, a pooling layer, and a fully connected layer. The number of these layers can vary according to the requirement of the task. These layers perform various operations on the data to identify unique features within the data. CNN is very useful for feature extraction which is exploited in this proposed approach.

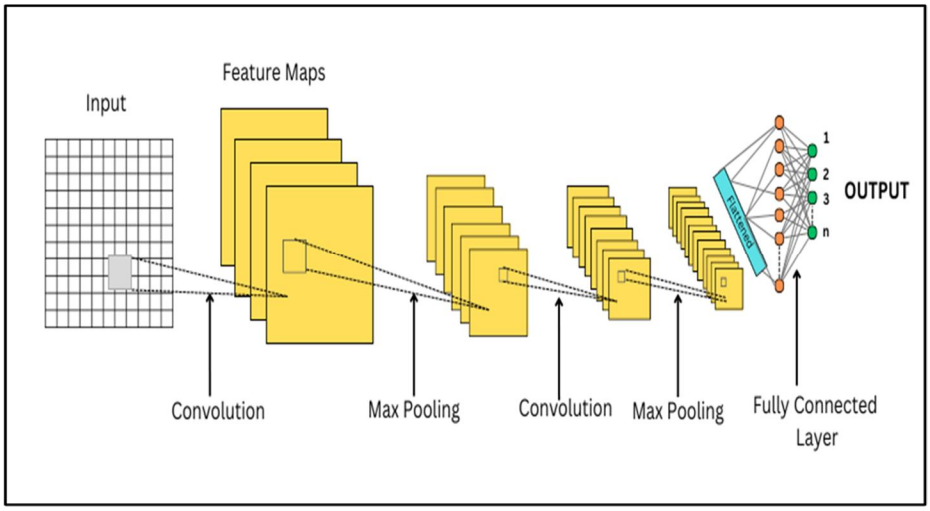


Fig. 3 Architecture of CNN

LSTM, (long short-term memory) networks, are a type of RNN(Recurrent neural network) [20]. Its main advantage over traditional RNNs is its ability to work with long sequential data, particularly in sequence prediction problems. Unlike traditional RNNs, LSTM does not lose information from previous iterations, with help of 4 gates LSTM manages to keep all the relevant previous information making it better suited for long sequential data, rather than just individual data points like images. This makes it well-suited for applications that involve sequential data such as machine translation. LSTM is a unique type of RNN that has demonstrated exceptional performance across a wide range of problems. LSTM architecture is illustrated in fig 4.

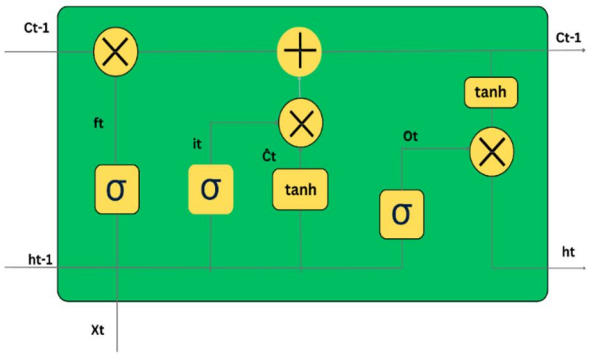


Fig. 4 . LSTM architecture

C. Attention Mechanism

In recognizing emotions from speech, the attention mechanism can be employed to direct the model's attention toward the portions of the audio signal that are most effective in conveying emotional information [21]. This can help to filter out irrelevant information and improve the overall performance of the model. The core goal of attention is to locate and assign distinct weights to distinct sections of the input sequence according to their relevance to accurate output. This technique is being employed in numerous fields like machine translation processing and computer vision applications.

Attention can be implemented in different ways, but the basic idea is to compute a weight or attention score for each element in the input sequence based on its relevance to the current output element. The attention weights are then used to compute a context vector, which is a weighted sum of the input elements. The context vector is then concatenated with the output element and fed into the next step of the model. There are several types of attention mechanisms, including additive attention, multiplicative attention, and self-attention. Additive attention computes the attention weights as a linear combination of a learned parameter matrix and the current output element. Multiplicative attention computes the attention weights as the dot product between the learned parameter matrix and the current output element. Self-attention calculates the attention weights by considering the similarity between different positions in the input sequence. The use of the Attention mechanism has improved the performance of deep learning models in a variety of tasks, particularly in cases where the input data is long or complex

D. Model Architecture

The model takes as input a sequence of features obtained from raw audio, i.e... MFCCs and uses a combination of convolutional and LSTM layers to obtain relevant features from the audio signal. The self-attention mechanism is then applied to the output of the LSTM to allow the model to focus on the most important parts of the audio sequence for identifying the emotional content.

Input layer: The input layer takes as input a sequence of audio features, such as Mel-frequency cepstral coefficients (MFCCs), which have been preprocessed from the raw audio signal. The 4 convolutional layers apply a series of filters to the input sequence to extract relevant features. Each filter is applied across the time dimension of the input sequence, with the output of each filter fed into the next layer. The number and size of filters, as well as the padding and stride used, are tuned to optimize performance.

The LSTM layers process the output of the convolutional layers to capture the temporal dependencies in the input sequence. Each LSTM layer consists of several LSTM units, which have internal gates that control the flow of information through the layer. The number of LSTM layers and units are tuned to optimize performance. The self-attention mechanism is applied to the output of the LSTM layers to allow the model to concentrate on the most important parts of the audio sequence for identifying the emotional content. This is attained by computing a weighted sum output of the LSTM at each time step, where the weights are learned based on the importance of each time step for identifying the emotional content. Fully connected layers: The output of the self-attention mechanism is typically fed into one or more fully connected layers, which transform the output into a vector of probabilities for each emotion class.

Output layer: The output layer uses a SoftMax activation function to convert the vector of probabilities. The class of emotion with the maximum probability percentage is selected and given as the output emotion for the given audio sequence. Model architecture is illustrated in Fig. 5

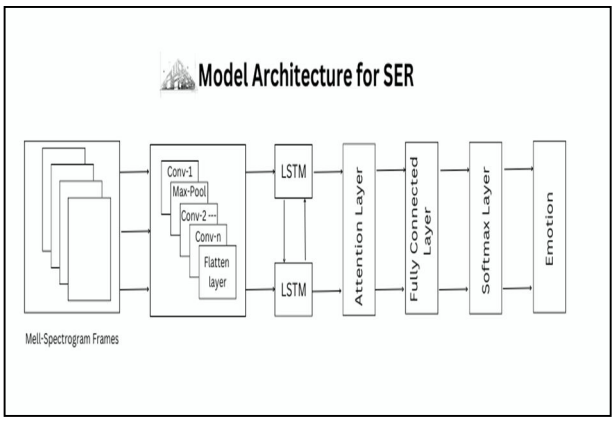


Fig. 5 Model Architecture

IV. EXPERIMENTAL RESULTS

A. Dataset

We will use audio-only files in the form of 16-bit, 48kHz .wav format, obtained from the RAVDESS dataset. The complete dataset comprises audio and video files of speech and song, which can be accessed from Zenodo and has a size of 24.8 GB. The RAVDESS dataset includes 1440 audio files, which consist of 60 trials per actor for a total of 24 actors. The equal gender split of male and females, deliver 2 same sentences in an American tone while expressing seven different emotions in 2 different intensities. The construction and perceptual validation of the RAVDESS dataset is explained in an Open Access paper published in PLoS ONE.[22]

B. Results and Discussions

The model was implemented using Python 3.3.8 and relied on several libraries, including TensorFlow 2.4.0, NumPy 1.19.5, Pandas 1.2.4, and Librosa 0.9.1. The tests happened on a Windows 10 OS, Intel(R)Core (TM) i-7 CPU @ 3.00 GHz processor, and 16-GB memory. To train the model, 32 samples were used in each batch, and the model was run for 200 epochs. The loss function applied was “categorical cross-entropy”, while the optimization function utilized was Adams with a learning rate (lr) of lr = 0.0001. Early stopping was implemented to prevent overfitting. The model takes a sequence of 30 audio features, namely MFCCs, as input. The model architecture consists of 4 2D CNN layers with Rectified Linear Unit (ReLU) activation functions and Long Short-Term Memory (LSTM) recurrent layers with 128 memory cells for learning the temporal aggregation. During training, a 20% dropout was applied to all layers to prevent overfitting. The model achieved an accuracy of 89%.

	Precision	Recall	F1-score	Support
Angry	0.94	0.89	0.91	36
Disgust	0.94	0.94	0.94	33
Fear	0.83	0.86	0.84	28
Happy	0.88	0.92	0.90	50
Neutral	0.70	1.00	0.83	19
Sad	0.86	0.83	0.84	46
Surprise	1.00	0.91	0.95	33
Calm	0.90	0.81	0.85	43
Accuracy			0.89	288
Weighted average	0.89	0.89	0.89	288

Fig. 6 Results for each emotion

Fig. 6 presents a statistical analysis of the proposed model's results for each emotion category, including accuracy and F1-score.

Fig. 7 Confusion Matrix

The confusion matrix of the proposed model testing, indicating the numerical representation of the emotion categories is displayed in Figure 7.

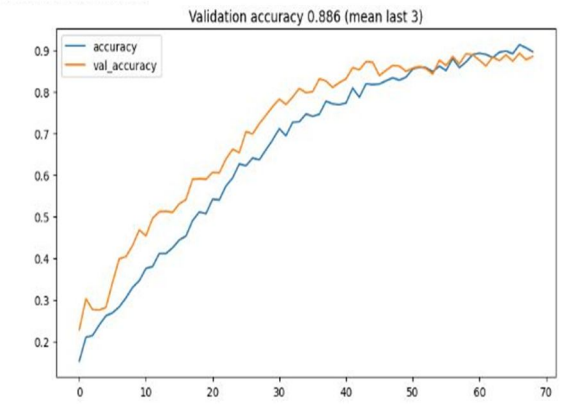


Fig. 8 Model Accuracy with respect to epoch

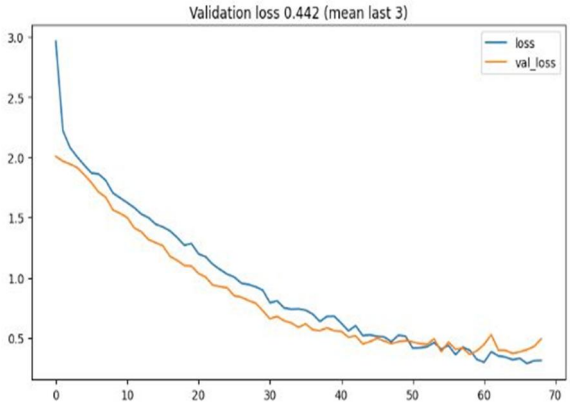


Fig. 9 Model loss with respect to epoch

The train versus validation accuracies and loss of the proposed model over the 200 epochs are depicted in Figures 8 and 9, demonstrating a stable training process for each epoch.

TABLE I
MODEL PERFORMANCE EVALUATION

Paper	Model Used	Accuracy
The Effects of Normalization Methods on Speech Emotion Recognition	CNN + LSTM	72%
Speech Emotion Recognition Using a Deep Neural Network	DNN	68.51%
Speech Emotion Recognition Using ANN on MFCC Features	ANN + MFCC	88.72%
Proposed Model	CNN + LSTM + Attention	89.83%

Table I compares our proposed model with other models, considering the methodology used and accuracy. Our model outperforms existing SER models.

V. CONCLUSIONS

This paper proposes a model for speech emotion recognition using an attention mechanism. The proposed model achieves a high accuracy of 89% for the 2D CNN-LSTM model with self-attention. The attention mechanism has been found to improve recognition performance by focusing on relevant emotional information and ignoring irrelevant information. This research has potential applications in clinical settings for the detection and treatment of mood disorders. The proposed model can assist physicians in understanding the emotional space of their patients and provide personalized care and medical service. Speech emotion recognition using deep learning has immense potential in clinical applications. With the development of more advanced deep learning models and large standardized datasets, the accuracy and generalizability of the models can be further improved, making it a valuable tool for healthcare and other related industries.

Future research could also explore the use of speech emotion recognition in other areas of medicine, such as detecting pain levels or predicting patient outcomes. Additionally, further studies could investigate the use of speech-emotion recognition in virtual therapy sessions, where patients can receive treatment remotely and still receive personalized care. Overall, the future of speech emotion recognition in clinical applications looks promising and has the potential to revolutionize mental healthcare

VI. ACKNOWLEDGMENT

It is our pleasure to present the research on “Attention-based Speech Emotion Recognition approach for medical application”. First, we would like to express our sincere gratitude to Mr. Ganesh S. Pise, Department of Information Technology for his invaluable guidance throughout this research project. His insights, expertise, and constant encouragement have been instrumental in shaping our ideas and helping us achieve our research goals. We are grateful for his patience, support, and mentorship, which have been critical to the successful completion of this research.

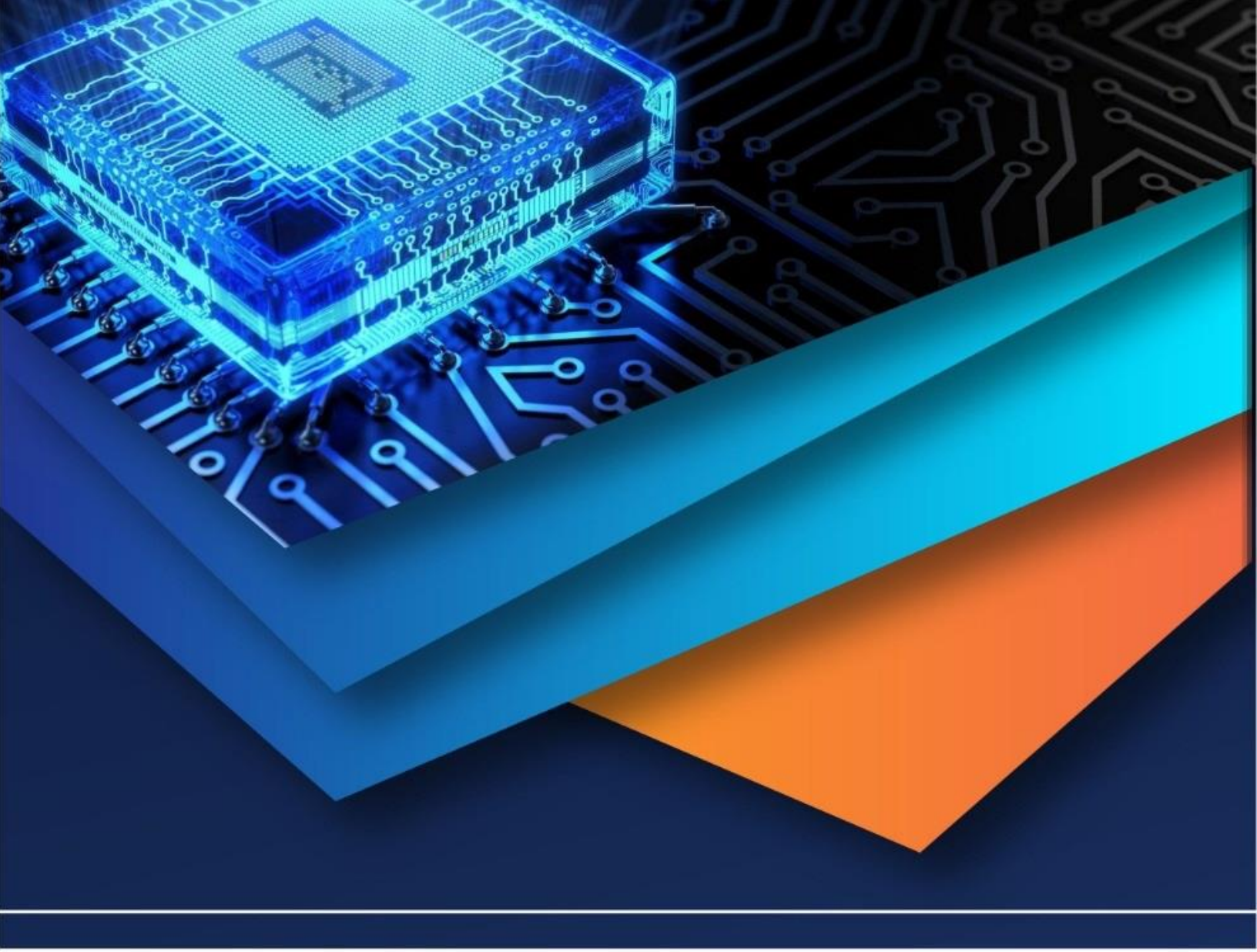
We also would like to thank Mrs. S. T. Kembhavi, Mrs. N. N. Shinde, and Mrs. R. D. Kapadi for reviewing our work and providing guidance.

We would also genuinely express our gratitude towards Mrs. R. V. Kulkarni, Project coordinator, Dr. A. S. Ghotkar, Head of the Department, and Dr. S. T. Gandhe, Principle for their kind help and cooperation.

We are thankful to all faculty members for their constant encouragement and assistance. And finally, we would like to thank our classmates for providing moral support and valuable suggestions

REFERENCES

[1] Patel, V. (2014). "Why mental health matters to global health". Transcultural Psychiatry, 51, 777 - 789. 2014
[2] T. H. Zhou, G. L. Hu, and L. Wang. "Psychological disorder identifying method based on emotion perception over social networks". International journal of environmental research and public health,16(6):953. 2019.
[3] T. W. H. O. (WHO) (2023). Suicide data. Accessed Jan 06, 2023, from https://www.who.int/teams/ mental-health-and-substance-use/data-research/suicide-data.
[4] The Washington Post (2022). There have been over 200 mass shootings so far in 2022. Accessed Jan 06, 2023, from https://www.washingtonpost.com/nation/2022/06/02/ mass-shootings-in-2022.
[5] P. Lieberman. The evolution of human speech: Its anatomical and neural bases. Current anthropology, Lieberman, P. (2007). The Evolution of Human Speech. Current Anthropology, 48, 39 - 66. 2007.
[6] B. Schuller, G. Rigoll, and M. Lang. "Hidden Markov model-based speech emotion recognition". IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03., 2, II-1. , 2003.
[7] K. Han, D. Yu, and I. Tashev. "Speech emotion recognition using deep neural network and extreme learning machine". Interspeech, 2014.
[8] S. K. Bhakre, and A. Bang. "Emotion recognition on the basis of audio signal using Naive Bayes classifier". International Conference on Advances in Computing, Communications and Informatics (ICACCI) 2363-2367, 2016.
[9] S. Mirsamadi, E. Barsoum, and C. Zhang. "Automatic speech emotion recognition using recurrent neural networks with local attention". IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2227-2231. 2017.
[10] F. Noroozi, T. Sapiński, & D. Kamińska. "Vocal-based emotion recognition using random forests and decision tree". International Journal of Speech Technology, 239-246, 2017.
[11] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik. "Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network". International Conference on Platform Technology and Service (PlatCon), 1-5, 2017.
[12] H. S. Kumbhar, and S. U. Bhandari. "Speech Emotion Recognition using MFCC features and LSTM network". IEEE International Conference on Computing, Communication and Automation (ICCCA), 1-3 2019.
[13] Mustaqeem, M. Sajjad, and S. Kwon. "Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM", IEEE Access, 8, 79861-79875, 2020.
[14] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, H. Mansor, M. Kartiwi , and N. Ismail. "Speech Emotion Recognition using Convolution Neural Networks and Deep Stride Convolutional Neural Networks". International Conference on Wireless and Telematics (ICWT), 1-6, 2020.
[15] T. Bänziger, & K. Scherer. "The role of intonation in emotional expressions". Speech Communication, 46(3-4), 252-267, 2005.
[16] A. Tsanas, A. M. Little, P. E. McSharry, and L. O. Ramig. "Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests". IEEE Transactions on Biomedical Engineering, 58(4), 884-893, 2011.
[17] G. Saposnik, R. Teasell, M. Mamdani, J. Hall, W. McIlroy, D. Cheung, and M. Bayley. "Effectiveness of virtual reality using Wii gaming technology in stroke rehabilitation: A pilot randomized clinical trial and proof of principle". Stroke ;41(7):1477-84, 2016.
[18] M. J. Bovin, E. J. Wolf, P. A. Resick, and B. P. Marx. "Using interpretable machine learning models to improve PTSD diagnosis". Journal of Anxiety Disorders, 42, 62-72, 2016.
[19] S. Davis, and P. Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences". IEEE Trans. Acoust. Speech Signal Process, 1980.
[20] S. R. Livingstone, and F. A. Russo. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English". PLoS ONE 13(5): e0196391. 2018..



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY