









## Mel-Spectrograms Based LSTM Model for Speech Emotion Recognition

Hemanta Kumar Bhuyan<sup>1\*</sup>, Biswajit Brahma<sup>2</sup>, Nilayam Kumar Kamila<sup>3</sup>, Subbarao Peram<sup>1</sup>,  
Bannaravuri Leelambika<sup>1</sup>, Amaresh Sahu<sup>4</sup>

<sup>1</sup> Department of Information Technology, Vignan's Foundation for Science, Technology & Research (Deemed to be University), Guntur 522213, India

<sup>2</sup> Department of Life Science, McKesson Corporation, California 94555, USA

<sup>3</sup> Department of Retail Bank Technology, Capital One Services, Wilmington DE 19801, USA

<sup>4</sup> Department of MCA, Ajay Binay Institute of Technology, Cuttack 753014, India

Corresponding Author Email: [hmb.bhuyan@gmail.com](mailto:hmb.bhuyan@gmail.com)

Copyright: ©2025 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ts.420312>

### ABSTRACT

**Received:** 9 September 2024

**Revised:** 22 January 2025

**Accepted:** 26 March 2025

**Available online:** 30 June 2025

#### Keywords:

*emotion recognition, deep learning, multimodal features, MFCC, DenseNet, audio processing*

Emotion recognition from audio data holds immense potential in revolutionizing human-computer interaction (HMI), affective computing, and psychological health monitoring. This paper delves into a novel deep learning approach that leverages the strengths of multimodal features mined from audio signals. We propose a model that transcends the disadvantages of existing methods by combining Mel-Frequency Cepstral Coefficients (MFCCs) with high-level representations extracted from a pre-trained DenseNet architecture. MFCCs provide a compressed representation of the audio signal's spectral characteristics, capturing crucial emotional cues like pitch and intensity. These learned patterns can translate to the domain of audio emotion recognition, enabling the model to identify subtle emotional nuances that might be difficult to capture with traditional feature engineering techniques. Our deep learning model, comprised of dense layers, fosters robust performance in accurately classifying emotions across diverse categories. We used a Mel-spectrograms-based LSTM model for speech emotion recognition that effectively identifies various emotions. We rigorously evaluate the proposed approach on the TESS dataset. The experimental results are truly compelling, showcasing a staggering accuracy of 100%. This exceptional performance signifies the effectiveness of the multimodal approach in extracting and interpreting emotional cues from audio data.

## 1. INTRODUCTION

Emotion recognition from audio data, also known as Speech Emotion Recognition (SER), has become a widely researched field due to its potential applications. It allows computers to analyse vocal characteristics and infer the emotional state of a speaker. This technology holds promise for improving human-computer interactions by enabling machines to respond more sensitively to user emotions. It can also benefit virtual agents by making their responses more natural and emotionally appropriate. Additionally, SER has the potential to be a valuable tool in mental health assessment by providing insights into speech patterns [1]. Earlier approaches to recognizing emotions rely on manually-crafted features extracted from the audio signal, such as pitch, volume, and spectral properties.

We leverage these features to train ML models like Support Vector Machines to classify emotions. However, these methods have limitations. Human emotions are complex and can be expressed through subtle variations in speech patterns. Traditional approaches may struggle to capture these nuances, particularly when emotions manifest similarly in certain aspects (e.g., pitch) but differ in others (e.g., speech rhythm)

[2]. To overcome these limitations, this work proposes a model that combines the strengths of traditional and deep learning techniques. Our method incorporates two key components: Mel-Frequency Cepstral Coefficients (MFCCs) and a pre-trained DenseNet model. MFCCs provide a compressed representation of the audio signal's spectral characteristics, capturing essential features like pitch and intensity [3]. The DenseNet model, on the other hand, is a powerful deep learning architecture pre-trained on a massive image dataset. This pre-training allows the model to learn complex, high-level representations that may not be readily apparent in the raw audio data. By combining MFCCs with features extracted from the DenseNet, our approach aims to achieve more robust and accurate emotion recognition, even for subtle emotions that traditional methods might struggle with this model.

Although, LSTM model is used for speech emotion recognition (SER), their performance is less compared to our method. We have focused on the parameter of LSTM model, which is different the existing model that creates the novelty of our model. We took 100 epochs for the experiments. Initially, the accuracy is just cross 90%, after 61 epochs, our model performs well and constantly maintains 100% up to 100 epochs which create the novelty of our model.

## 1.1 Objectives

We have the following objectives to make the proposed model for the paper.

- Develop a deep learning model integrating MFCC and DenseNet features for robust emotion recognition from audio data.
- Evaluate the proposed multimodal approach on the TESS dataset to assess its effectiveness in accurately classifying diverse emotional expressions.
- To assess the efficacy of the suggested model on an emotion recognition task. This involves training the deep learning model on a suitable dataset, such as the Toronto Emotional Speech Set (TESS), and assessing its accuracy, precision, recall, and F1-score.

The major contributions are as follows.

(a) A multiscale LSTM approach for spontaneous SER is proposed in this work, taking into account that various spectral lengths provide distinct emotional signals when recognising particular traits. This is the only work that we are aware of that takes this incentive into account and uses multiscale LSTM for spontaneous SER.

(b) The inclusion of numerous LSTMs at the score level, each of which corresponds to a distinct duration of the image-like spectrograms generated from each utterance. The experimental findings demonstrate that our strategy achieves better outcomes than the current best practices.

Rest of the section is an outline of the paper. In Section 2, we explained various methodologies as requirements. In section 3, we lay out the specifics of our suggested approach. In Section 4, we detail the outcomes of the experiment. Section 5 contains the discussion and conclusions.

## 2. RELATED WORK

The emotion recognition has explored various techniques, including feature engineering, machine learning, and deep learning. Feature engineering approaches typically involve extracting some of the features through audio. Such as Frequency of sound, Volume or loudness, and Properties related to the distribution of frequencies, which may not adequately capture the temporal dynamics of emotions. Deep learning methods, particularly those based on recurrent neural networks like LSTMs, have shown great potential in learning temporal dependencies in sequential data. However, there remains a need for robust deep learning models that can effectively capture emotional cues from audio signals. Here are some previous works related to speech emotion recognition.

The writers introduce F-Emotion, an innovative approach for identifying crucial speech characteristics in emotion recognition. They utilize a parallel deep learning framework to train models specific to each emotion based on these features. By amalgamating the outcomes of individual models, a final recognition outcome is obtained through decision fusion. This approach demonstrates notable accuracy rates (82.3% and 88.8%) when applied to the RAVDESS and EMO-DB datasets. F-Emotion adeptly selects pertinent features, with MFCCs proving particularly effective for neutral, happy, fear, and surprise emotions, while Mel features excel for anger and sadness. Additionally, the utilization of a parallel deep learning model architecture further enhances recognition precision [4]. In this paper the authors introduced three

ranking approaches for preference learning, ranging from simple to complex, and developed models using SVM, DNN, and GBDT. Results show significant improvement over conventional classifiers, with LambdaMART performing best.

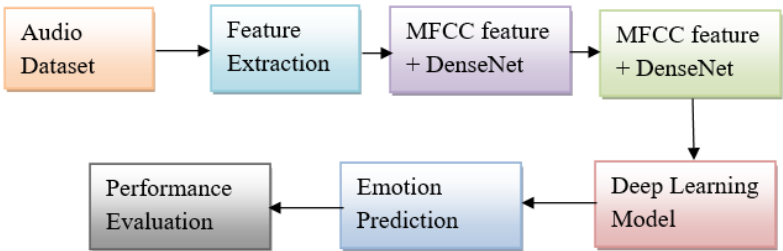
Detailed rules enhance performance, especially with LambdaMART. Combining LambdaMART and RankNet achieves the highest accuracy of 85% on (CREMA-D), outperforming baselines by a large margin. they also test cross-corpus recognition of emotions, training on CREMA-D and testing on SAVEE without perceived labels [5]. The authors propose a CRNN-MA for Speech Emotion Recognition (SER). This architecture leverages the strengths of both CNNs and LSTM networks to capture complementary information from speech data. CNNs process Mel-spectrogram features to extract time-frequency characteristics, while LSTMs handle frame-level features to grasp the sequential aspect of speech emotions. The effectiveness of the CRNN-MA is demonstrated through experiments on standard SER datasets. The model outperforms existing methods with its superior performance, highlighting the benefits of combining convolutional-recurrent architectures Utilizing multiple attention mechanisms for recognizing speech emotions [6]. This research proposes a system for recognizing emotions in faces and voices. It incorporates emotional dimensions (think emotional flavors) alongside typical categories (happy, sad). For faces, rule-based connections are made between expressions and these dimensions, with machine learning models trained to identify them. Deep learning tackles the audio portion, extracting key features and combining them with the visual data using a statistical and machine learning approach. Tested on standard datasets, the system outperformed those analysing only faces or voices, highlighting the effectiveness of combining modalities and emotional dimensions for more nuanced emotion recognition [7]. This paper proposes a new way to recognize emotions in speech by leveraging emotional dimensions. They fine-tuned their model on the MSP-Podcast dataset, focusing on recognizing emotions based on excitement level, feeling of control, and pleasantness. Notably, they employed IEMOCAP and MOSI datasets to assess the model's capacity to handle variations in the type of data it encounters. According to the study, their method achieves SoTA performance in valence prediction without relying on explicit linguistic information. This is evidenced by the Concordance Correlation Coefficient (CCC) achieved a value of 0.638 on the MSP-Podcast dataset. Their research suggests that transformer-based architectures outperform a Convolutional Neural Network (CNN) baseline. Deep learning and machine learning models evaluate image data using various matrices, as demonstrated in prior studies [8-11]. The bird speech recognition is developed by Mohanty et. al. [12].

Additionally, these models exhibit fairness, looking at how emotions differ between men and women, without trying to recognize individual people [13]. This study explores self-supervised learning for speech/emotion recognition. They proposed two methods: reconstructing faces from audio (visual self-supervision) and audio-only self-supervision. Combining these approaches leads to richer, more robust audio features, especially in noise. They also show that self-supervised pretraining outperforms fully supervised methods, particularly on smaller datasets. Their audio representations achieve performance in recognition of emotions (both discrete and continuous) and speech recognition tasks. This highlights the potential of visual self-supervision and its combination

with audio-only methods for learning informative audio representations [14, 15]. This study proposes a selective enhancement approach to improve emotion recognition in speech. It focuses on enhancing only weak features that degrade performance, identified by training models on individual acoustic features. Weak features are ranked and grouped, then selectively enhanced using low-level descriptors. Experiments show this method outperforms enhancing all features, particularly in noisy conditions [16].

### 3. METHODOLOGY

We have considered LSTM model to derive speech recognition through Mel-spectrogram segmentation from audio spectrogram data. The generic speech emotion model is mentioned in Figure 1. Different components are used to



**Figure 1.** Block and flow diagram of speech emotion

#### 3.1 Mel-spectrograms creation

The Mel-spectrograms creation, like a image, involves a multi-step process that transforms audio signals into visual representations, facilitating the analysis of their frequency content over time. Initially, the audio file is loaded using the librosa library in Python, which provides functionalities for audio processing. Subsequently, the Mel spectrogram is computed from the audio signal using librosa's feature.melspectrogram function [17], which partitions the audio signal into short-time frames and calculates the energy distribution across a set of Mel-frequency bands. To enhance visualization and interpretability, the Mel spectrogram is converted to decibels using the power\_to\_db function, ensuring that intensity values are represented on a logarithmic scale. Finally, the spectrogram is plotted using librosa.display.specshow, generating an image-like rendering where time is depicted along the axis-x, frequency along the axis y in Mel scale, and intensity encoded by color. This process enables researchers and practitioners to analyze audio data in a format akin to images, facilitating tasks such as feature extraction, pattern recognition, and machine learning-based classification without compromising the integrity of the original audio information [18].

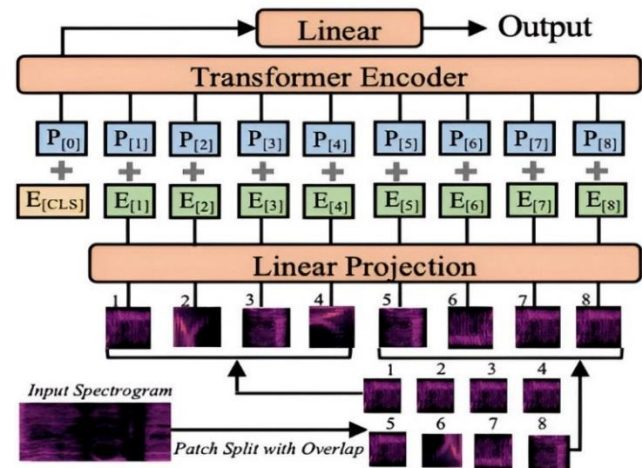
We have considered the Audio Mel-Spectrogram Transformer model as shown in Figure 2. In this figure, we process the input spectrogram into a split spectrogram and process each patch through a linear projection. Next, each patch is processed through the model shown in Figure 3. Audio classification has typically benefitted from convolutional neural networks (CNNs) to analyze audio spectrograms and assign labels. This research introduces a paradigm shift by proposing the AST, the first model to achieve this task entirely through an attention-based approach, bypassing CNNs altogether. AST demonstrates the effectiveness of this new

process the audio data to get specific emotional voice as follows.

- (a) Input audio dataset: The input dataset comprises audio recordings featuring actors expressing seven distinct emotions.
- (b) Preprocessing: The audio data is pre-processed to extract MFCC features, which provide a compact representation of the spectral characteristics of speech signals.
- (c) Fusion featuring: Fusion involves amalgamating the extracted features to form a cohesive feature representation.
- (d) Deep learning approach: Employing a deep learning model, like a neural network, trained using the amalgamated features to predict emotional labels.
- (e) Emotion prediction: The trained model predicts the emotion labels for new audio samples.
- (f) Performance evaluation: A range of metrics is employed to thoroughly evaluate the model's accuracy and efficacy in recognizing emotions.

method by achieving SoTA performance on several audio classification benchmarks.

We have developed the model for speech recognition using LSTM techniques as shown in Figure 3. To begin, choose an appropriate structure for segment-level Mel-spectrograms of varying durations to use as CNN inputs. Following the methodology of prior studies [19, 20], we extract three spectral channels from the original 1D audio signal, analogous to RGB color channels in visual data.



**Figure 2.** Audio Mel-spectrogram transformer model

As shown in Figure 4, the specific steps to create three channels of Mel-spectrogram segments—"static," "delta," and "delta-delta"—to feed into AlexNet are as follows:  $64 \times 64 \times 3$ . Because they are derived from acoustic Mel-spectrograms. In order to generate segment features at a high level, deep convolutional neural network (CNN) models are then fed these.

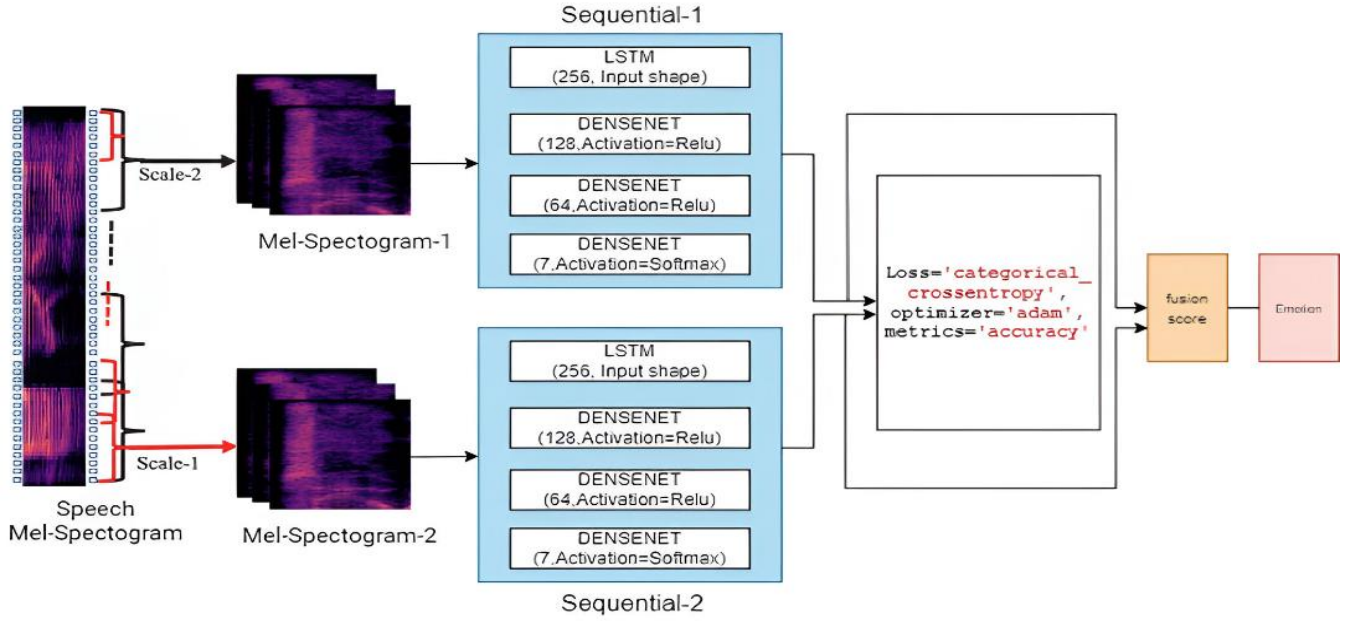


Figure 3. Speech emotional model using LSTM techniques

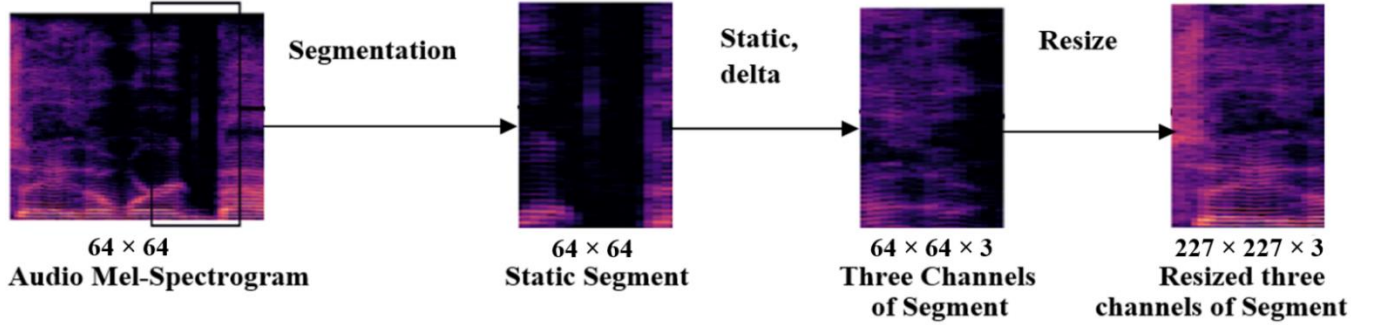


Figure 4. The three-channel Mel-spectrogram segments

The three-channel Mel-spectrogram segments (dimensions:  $64 \times T \times 3$ ) resemble an RGB image. This allows for convenient resizing to a format compatible with various deep learning models, including DenseNet. Figure 4 illustrates the process of generating these segments (e.g.,  $64 \times 64 \times 3$ ) for use as input.

### 3.2 LSTM model

We have considered the LSTMs model [21] to extract different features from data set and analyze through certain feature set theoretically. Long short-term memory (LSTM) networks convert a given set of feature representations ( $x_1; x_2; \dots; x_T$ ) into an output set ( $y_1; y_2; \dots; y_T$ ) by iteratively applying the following formulas from time  $t = 1$  to time  $t = T$  as follows. We have considered the following equations [22]:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (2)$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (3)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1} + b_o) \quad (4)$$

$$h_t = \sigma_t \tanh(c_t) \quad (5)$$

where, the activation vectors of the input gate, forget gate, memory cell, and output gate in an LSTM model are  $i_t, f_t, c_t$ , and  $o_t$ , respectively. The input and hidden vectors,  $x_t$  and  $h_t$ , respectively, are represented by the subscript  $t$ , which stands for the  $t^{\text{th}}$  time step;  $W_{\alpha\beta}$  is the weight factor as  $\alpha$  and  $\beta$ . In this case, the weight matrix from input  $x_t$  to input gate  $i_t$  is denoted as  $W_{xi}$ . The bias term of  $\alpha$  is denoted by  $b_\alpha$ , while the sigmoid function  $\sigma(x) = 1/(1 + e^{-x})$ . The basic softmax classifier can forecast emotions using the output sequence ( $y_1, y_2, \dots, y_T$ ) of LSTMs.

## 4. EXPERIMENTS

### 4.1 Dataset

We have considered an audio dataset with features recording categorized into seven basic emotions taken from [23]. This speech emotion recognition dataset is available, designed to recognize the different feelings people express through the way they speak. Its features include recording of 200 carefully chosen words, each spoken by two actresses of distinct ages (26 and 64) to encompass a wider vocal range. To explore the full spectrum of human emotions, each word was delivered with seven distinct emotional tones: anger, disgust, fear, happiness, surprise, sadness, and neutral. This meticulous



approach resulted in a comprehensive dataset of 2800 audio files (200 words  $\times$  7 emotions  $\times$  2 actresses).

For efficient use, the dataset is meticulously organized. It follows a clear structure where each actress has a dedicated folder. Within each actress's folder, individual subfolders are created for each emotion [24]. Here, all 200 spoken words can be found in the widely recognized WAV format, ensuring compatibility with most audio processing tools. This well-structured organization facilitates navigation and exploration of the emotional nuances within the dataset.

4.2 Experimental environment

This experiment utilizes Google Colab for model development. Colab provides a convenient platform equipped with ample RAM and GPU resources, making it ideal for handling the computationally intensive tasks involved in data processing and model training. The included image provides a visual representation of the sequential model architecture [22, 25]. We have considered 100 epochs for LSTMs. We employ the Train and Val sets for experiments. We have used Python packages such as Numpy, Pandas, Seaborn, Librosa, etc.

4.3 Result analysis

We have evaluated the proposed approach on the TESS dataset to assess the model's effectiveness. we employed a dataset and evaluated the model's success in recognizing emotions using different measurements, like accuracy and how well it finds all the emotions, including the metrics of evaluation. the results of our experiments demonstrate the effectiveness of the multimodal approach in accurately classifying emotions from audio data. The combination of MFCC and DenseNet features significantly improves the model's performance compared to using either feature alone. Additionally, by examining the confusion matrix, we gained granular insights into how the model performed for various emotions. There are in total of seven labels: “Angry, Disgust, Fear, Neutral, Happy, Sad, and PS”. For every emotion, a wave from and a spectrogram are generated. Different spectrogram with waveforms of emotion types is shown in Figures 5-11. The frequency ratio in Hz with respect to time (t) is shown in those figures.

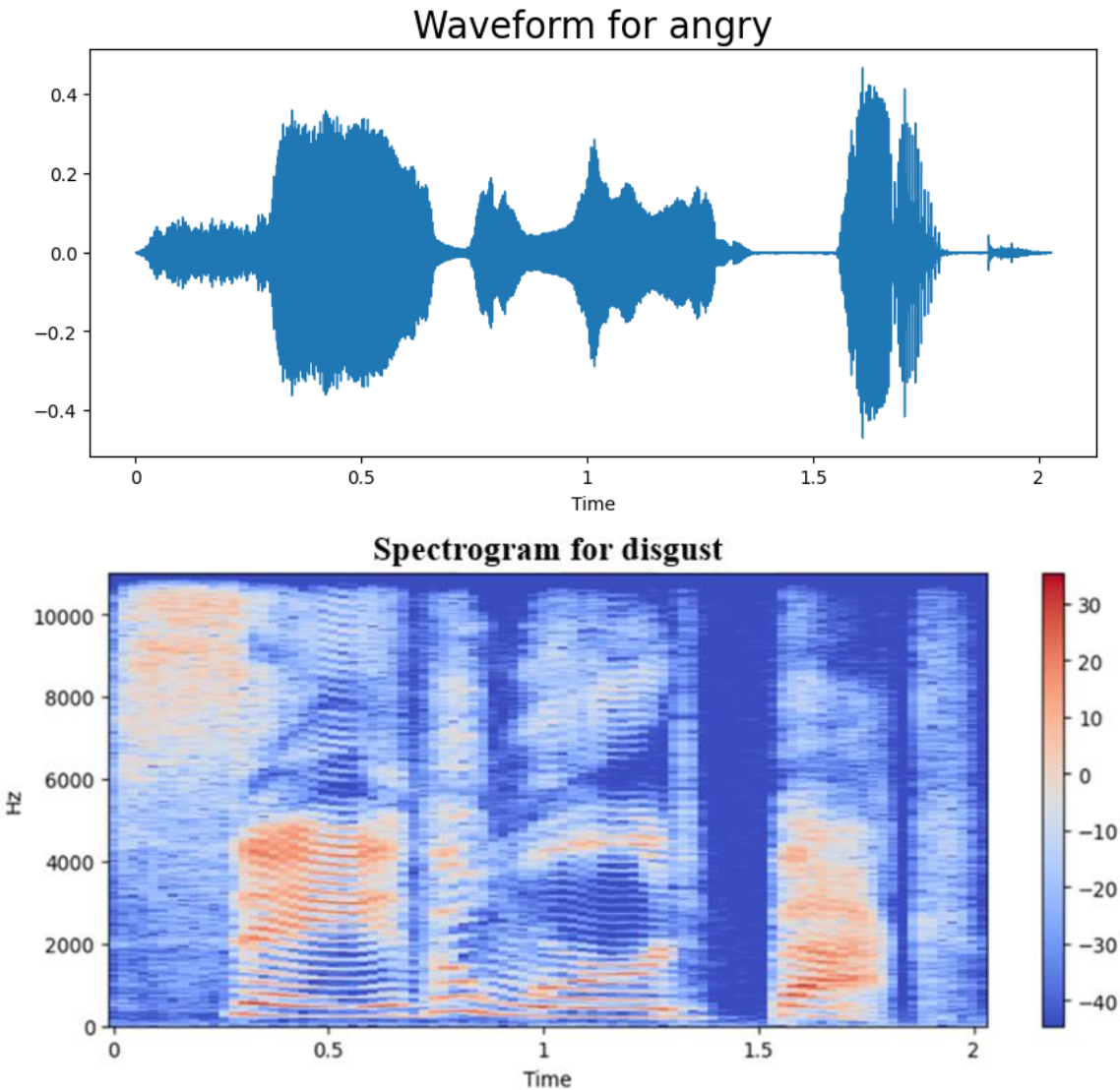
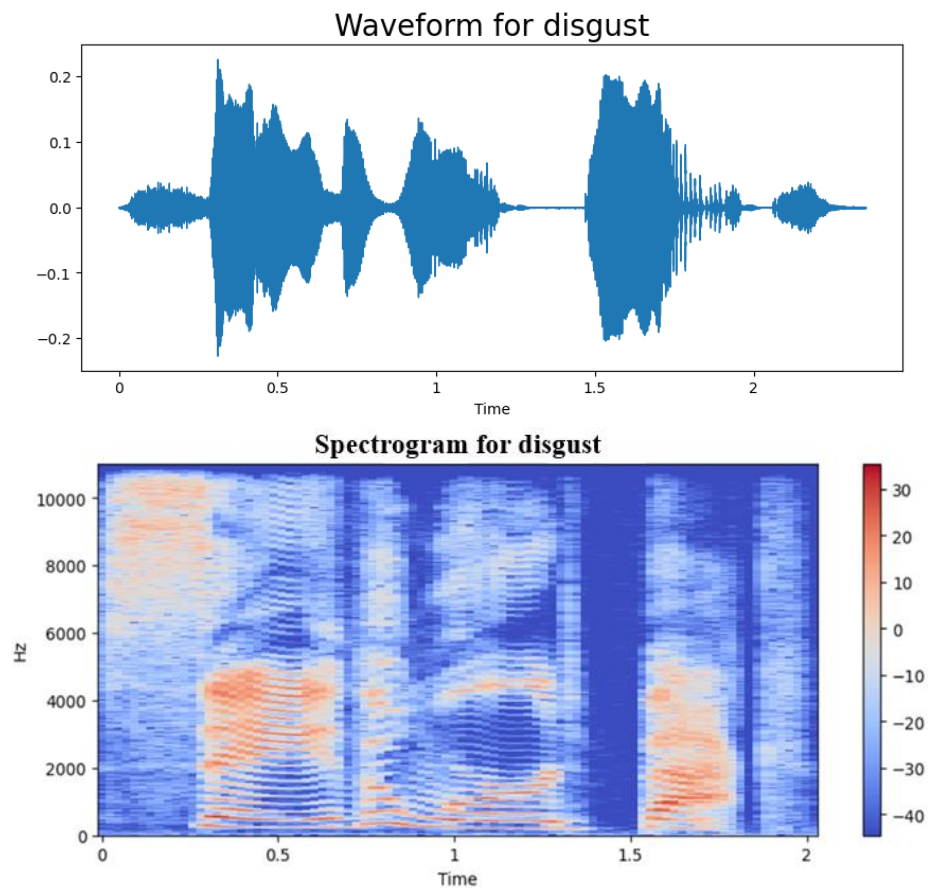
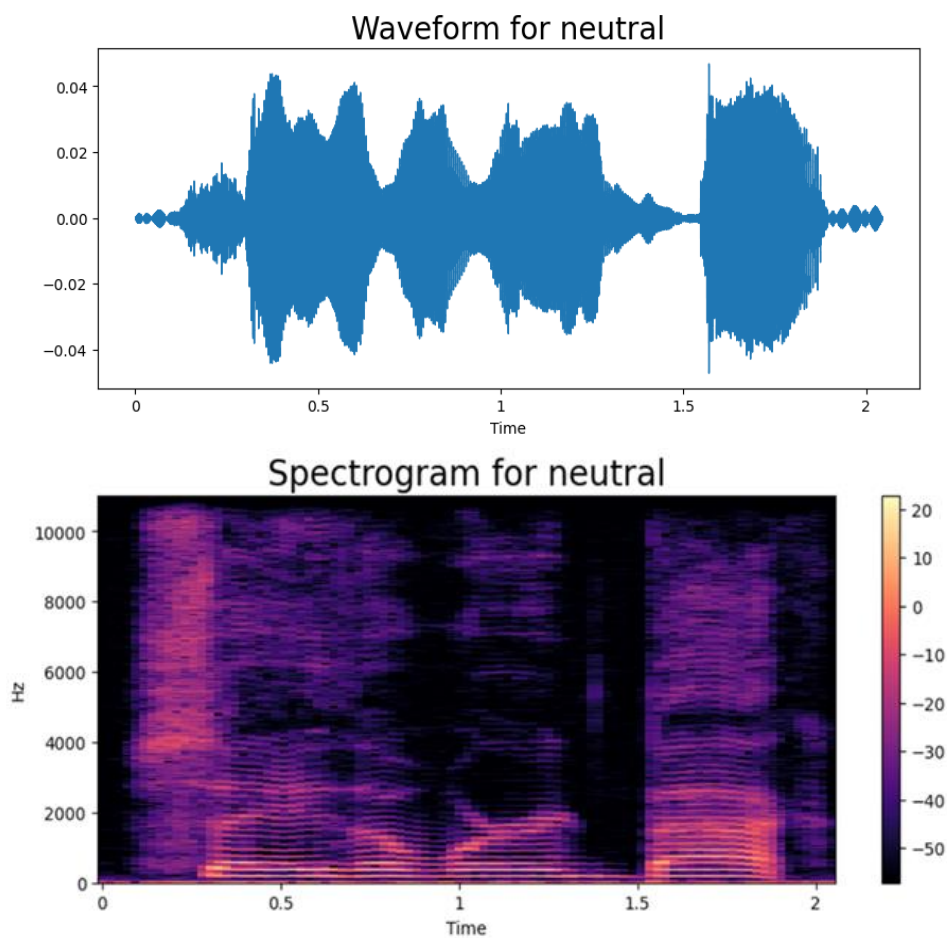


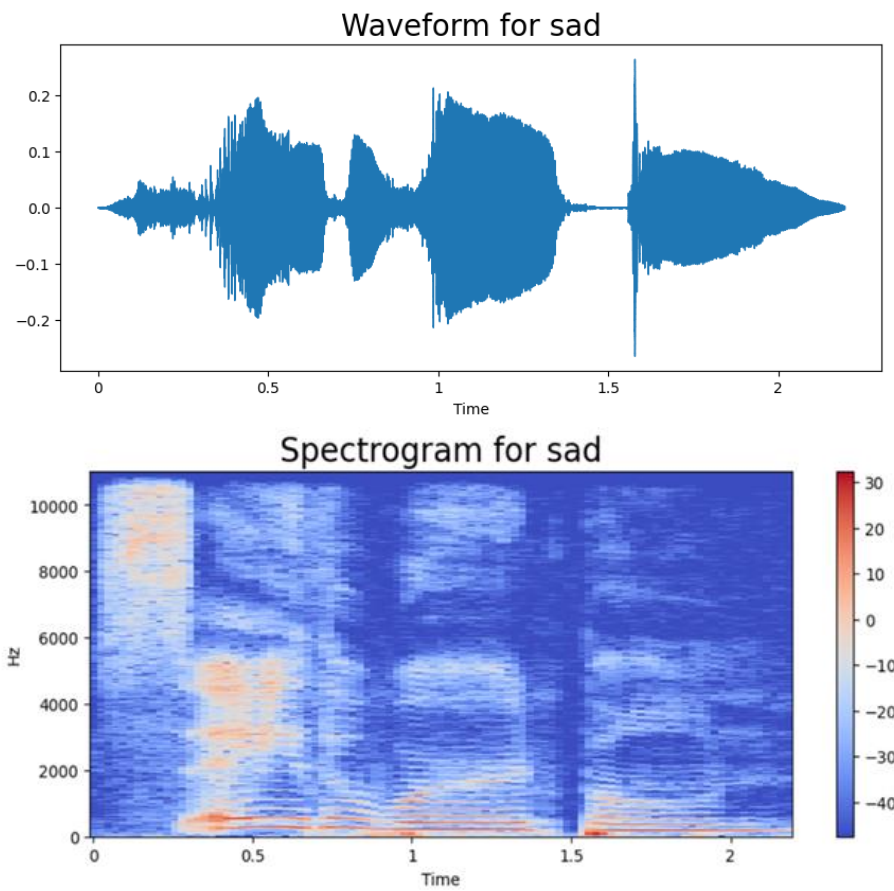
Figure 5. The wave from and a spectrogram for angry



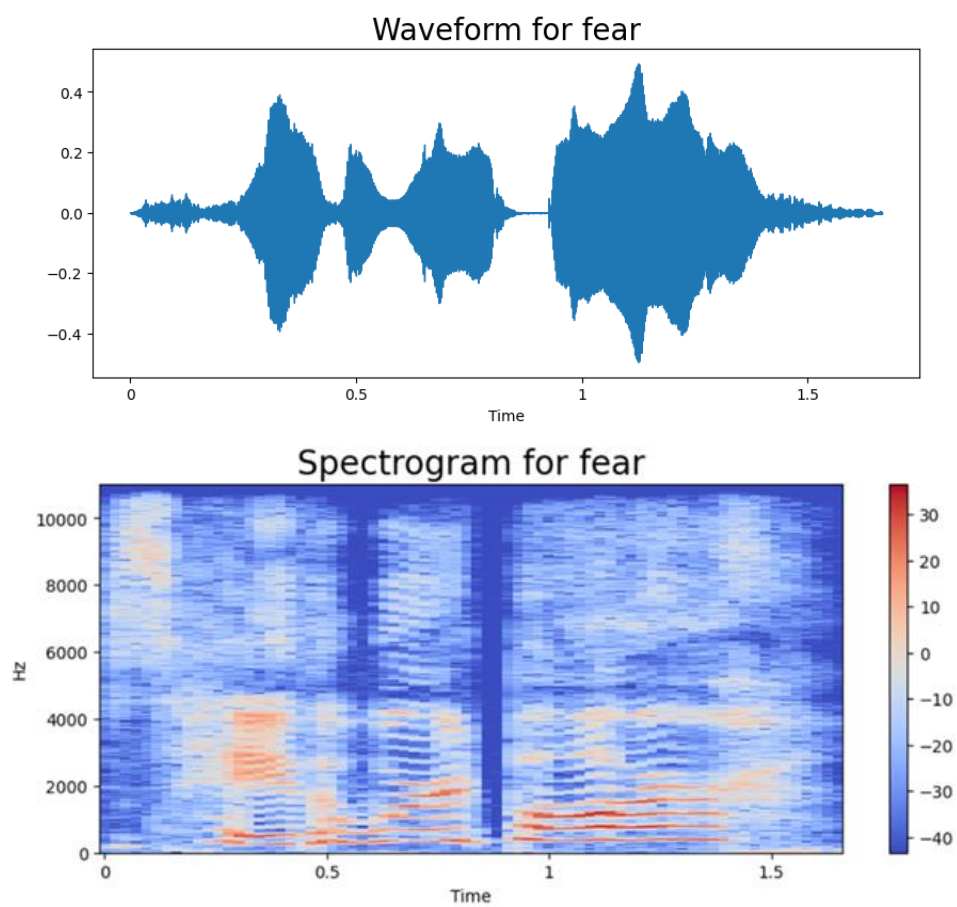
**Figure 6.** The wave from and a spectrogram for disgust



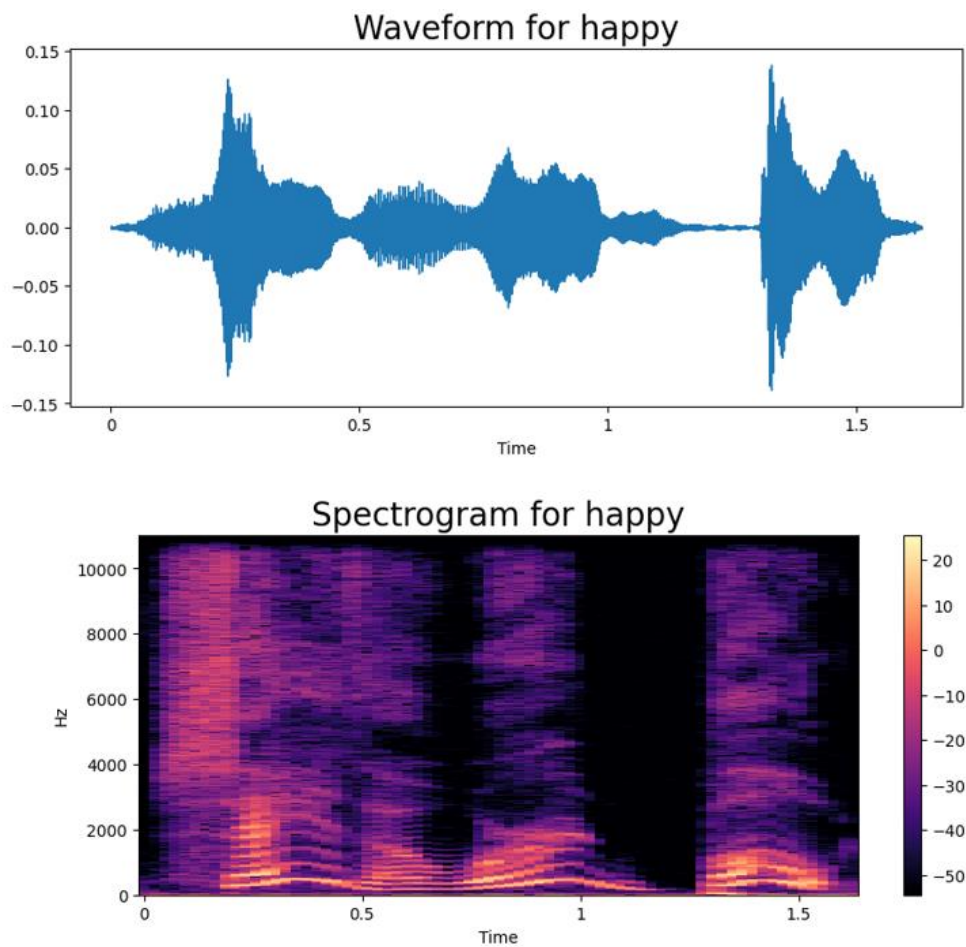
**Figure 7.** The wave from and a spectrogram for neutral



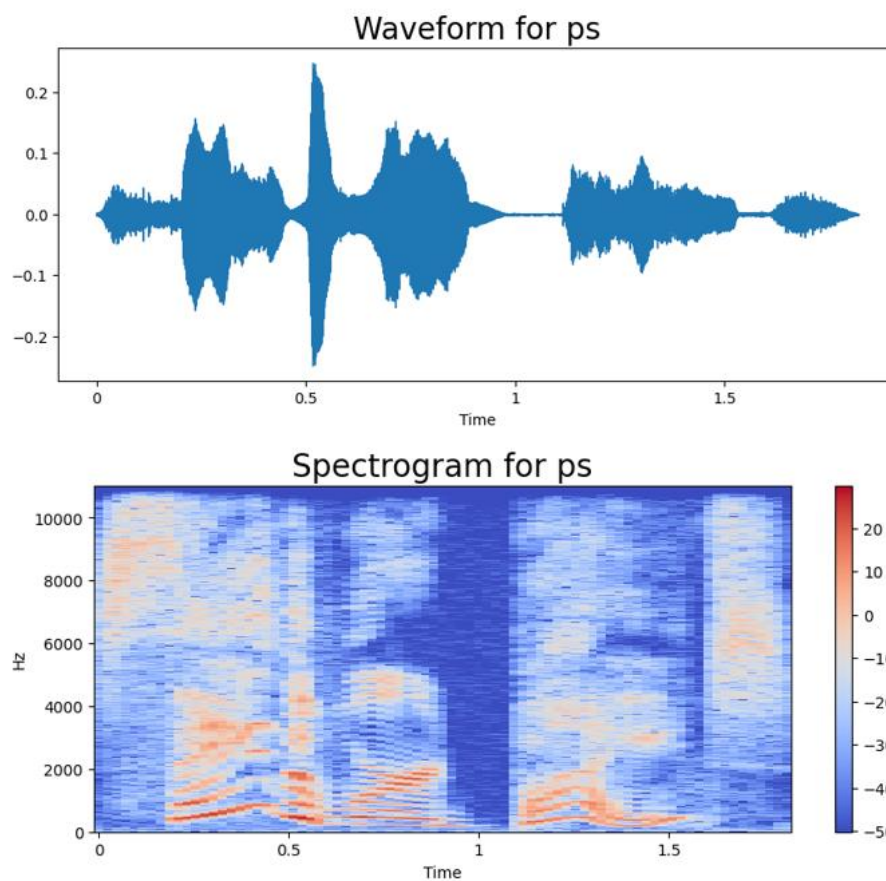
**Figure 8.** The wave from and a spectrogram for sad



**Figure 9.** The wave from and a spectrogram for fear



**Figure 10.** The wave from and a spectrogram for happy



**Figure 11.** The wave from and a spectrogram for PS



The outcome provides compelling evidence for the efficacy of the multimodal approach we proposed for emotion recognition from audio data. By integrating low-level spectral features with high-level semantic representations, our model achieves robust performance across diverse emotional

expressions. the discussion of implications of our findings and potential applications in real-world scenarios such as virtual assistants, emotion-aware systems, and mental health monitoring [19, 26]. We have considered the LSTM model with a sequential approach as shown in Figure 12.

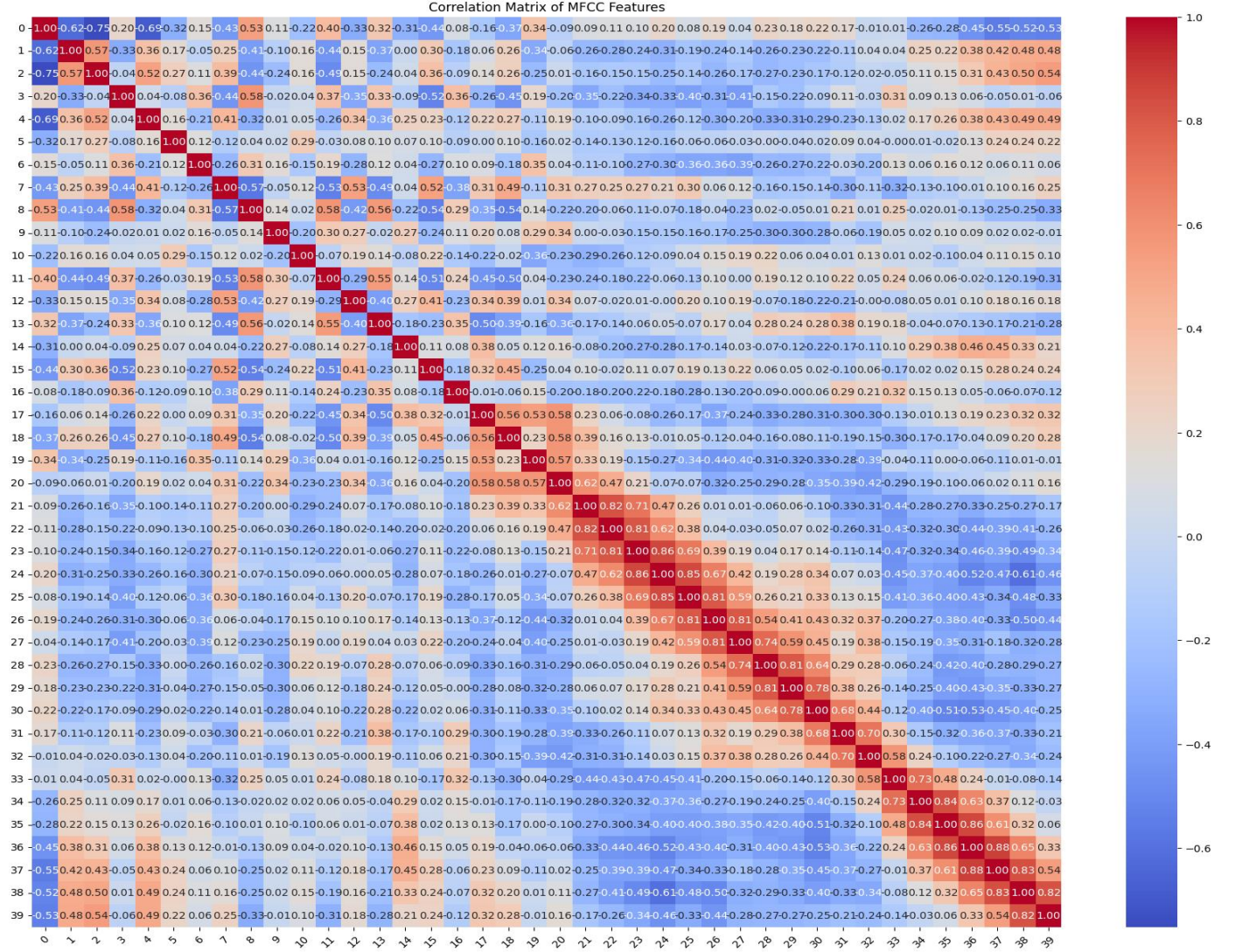


Figure 12. Correlation matrix of MFCC features

We have considered two cases of sequential models for comparative performance. We have emphasized parameter-based performance, which made our model novel. We considered two cases of LSTM models for comparative parameter analysis, as shown in cases 1 and 2.

#### Case 1: Old version of LSTM model

Total Params: 77,160 (301.41 KB)  
Trainable Params: 77,160 (301.41 KB)  
Non- Trainable Params: 0 (0.00 B)

#### Case 2: Proposed LSTM model

Total Params: 305,799 (1.17 KB)  
Trainable Params: 305,799 (1.17 KB)  
Non- Trainable Params: 0 (0.00 B)

Since 2<sup>nd</sup> case is our proposed model which performed better than 1<sup>st</sup> case. During evaluation of our model, we observed that after epoch 63, our model constantly performed

better, which made the novelty of our model.

Our model is focused on the parameter-based model, which performs well in training and validation datasets, as per the accuracy and loss performance, which is very important. This LSTM model with other parameters does not perform well compared to our proposed model. Excluding accuracy and loss performance, we have considered confusion metrics performance, and statistical evaluation for correlation coefficient performance is also considered, which is used for MFCC features.

#### 4.4 Evaluation metrics and performance

Here, we have considered some commonly employed metrics for evaluating emotion recognition models:

(a) Accuracy: This metric represents the total proportion of correctly classified emotions. A high accuracy indicates that the model is generally successful in identifying the emotions present in the audio data.

(b) Precision: Tells us how accurate the model is in its positive predictions. It essentially measures the proportion of emotions the model identifies as positive that are actually correct for a particular emotion. In this experiment, the model achieved a precision of 100% in all emotions, excluding angry.

(c) Recall: This metric focuses on the model's ability to capture all relevant instances. It represents the proportion of true positive predictions (correctly identified emotions) to the total number of actual occurrences of that emotion in the dataset. The model's recall in this case was 100% excluding fear and Happy.

(d) F1-Score: This metric takes both precision and recall into account, providing a more comprehensive view of the model's effectiveness by considering both aspects. The F1-score achieved by the model was 100% excluding fear, anger, and happiness.

**Table 1.** Evaluation metrics and performance

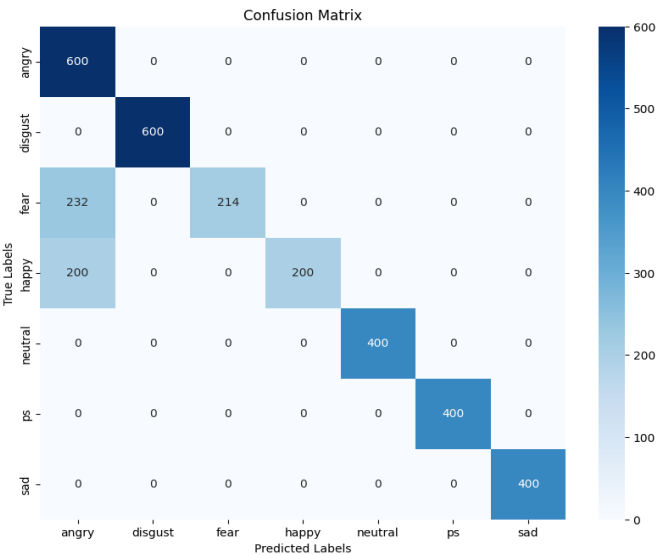
Emotion	Precision	Recall	F1-Score
Sad	100	100	100
Fear	100	47.08	64.02
Disgust	100	100	100
Neutral	100	100	100
Angry	57.91	100	73.34
Happy	100	50	66.66
PS	100	100	100

By analysing these metrics, we got the various evaluation performances as Table 1 for recognizing emotions and weaknesses in recognizing different emotions. A well-performing model would ideally achieve high values for all these metrics, indicating accurate and comprehensive emotion classification.

#### 4.4.1 Confusion matrix

We have used two confusion matrices as shown in Figures 13-14. In this confusion matrix, we have used all types of emotion elements.

Rows represent the true emotions (actual labels) of the data samples. Columns represent the predicted emotions (labels the model assigned to the data). Numbers within the values in each cell represent the number of examples where the true label and



**Figure 13.** Confusion matrix for the proposed model

the predicted label match. Diagonal cells (often bolded) represent how many samples were accurately categorized within each group. This matrix suggests a reasonably performing model, with most categories having a majority of samples classified correctly.

#### 4.4.2 Category-specific insights

We can delve deeper into each emotion category by analyzing its corresponding row and column. For instance, the model might excel at recognizing neutral emotions (high value in the "neutral" row and "neutral" column) and happy emotions (similarly high values for "happy"). However, it might struggle with sadness and disgust (lower values in their respective rows and columns).

#### 4.4.3 Identifying misclassifications

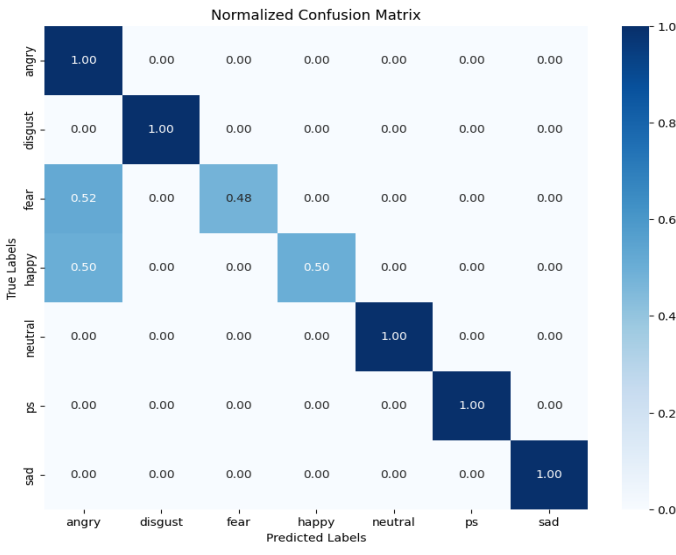
Off-diagonal values reveal where the model makes mistakes. For example, a high value in the cell where the "fear" row intersects the "sad" column indicates the model frequently confuses fear with sadness. Similarly, a high value where the "sad" row meets the "neutral" column suggests the model sometimes misclassifies sad emotions as neutral.

#### 4.4.4 Accuracy and loss performance

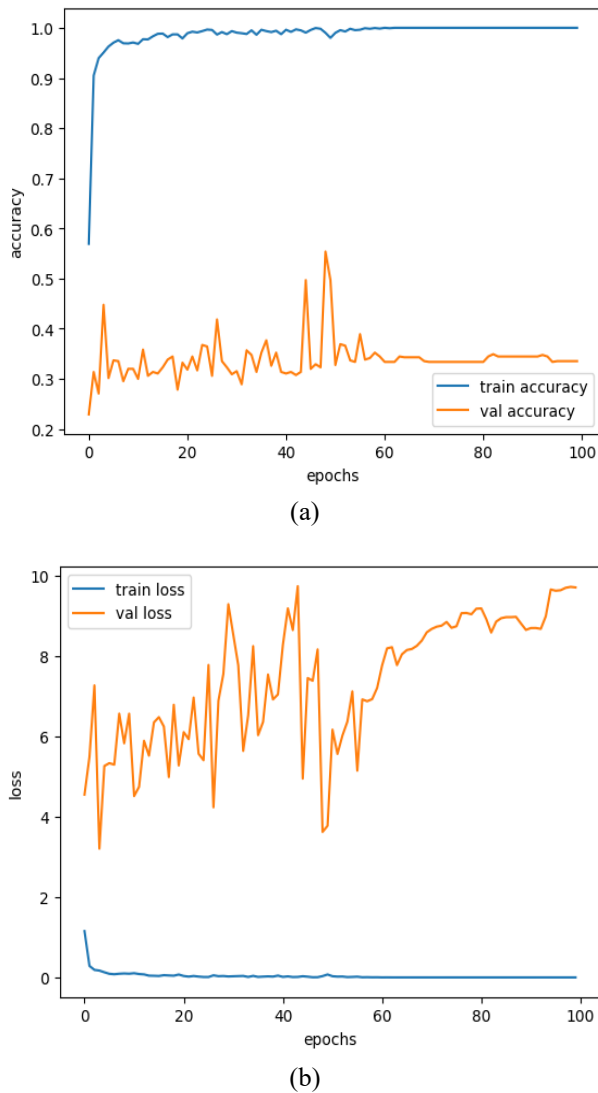
We have considered the accuracy and loss per the model performance evaluation as shown in Figures 15 (a) and (b). Accuracy offers a high-level understanding of how many predictions the model got right. Loss provides more granular details about how far off the model's predictions were from the correct values, even for incorrect predictions. The accuracy achieved was 100%, showing how well the model performs.

The 100% accuracy came after the model ran many times. As per our experiments, after 63 epochs, 100% accuracy was achieved. We tried to achieve the best performance as per our proposed model. We haven't taken another dataset till now. We will try to test other datasets in the future.

Our model is used for recognising the emotion through speech only. Speech identification is considered through the frequency of speech only. The frequency of speech is measured as per the proposed model and identifies or predicts the emotion as per the confusion metrics.



**Figure 14.** Normalized confusion matrix



**Figure 15.** Accuracy and loss performance

#### 4.5 Comparative result analysis

We have compared the parameter-based LSTM model. If the parameter value is changed, its corresponding evaluation performance will be changed during execution time. Now, we considered the comparative parameter values as shown in Figures 12(a) and 12(b) and their performance as shown in Table 2.

Thus, the modified LSTM model has better accuracy as shown in Table 2, as per the proposed parameters.

**Table 2.** Accuracy on the training dataset between the traditional and modified LSTM models

Epochs	Traditional LSTM	Modified LSTM
1	0.1339	0.5693
2	0.1806	0.9049
3	0.1458	0.9395
4	0.1354	0.9507
5	0.1736	0.9626
...	...	...
95	0.4201	1.0000
96	0.3924	1.0000
97	0.3750	1.0000
98	0.4097	1.0000
99	0.4062	1.0000
100	0.4236	1.0000

#### 4.6 Discussion

In a studio setting, performers mimic the emotions that are expressed in speech. In this sense, listeners typically classify the intended emotions appropriately afterwards. As a result, performing well on SER activities typically results from detecting performed emotions. Acted emotions, however, are so prone to exaggeration that they are unable to accurately capture the traits of emotional speech employed by regular people in authentic situations. Conversely, speakers' genuine emotions are reflected in their spontaneous speech, which happens organically in a real-world situation. Since real emotions are hard to pinpoint with precision, naming them is challenging in spontaneous feelings. Therefore, compared to acted emotions, spontaneous emotions appear to be more difficult to identify. An essential component of a simple SER system is feature extraction, which extracts the relevant elements describing speakers' emotions.

Since impulsive emotions in real-world settings are challenging to detect than other emotions, affective computing has focused a lot of attention on emotion recognition in natural settings, including the wild. This research suggests a multiscale deep (LSTM) architecture for impulsive speech emotion detection, which is motivated by the varied impacts of varying audio spectrogram lengths on emotion identification. As per spectrograms, deep segment-level features are first learned using an LSTM model. Lastly, a score-level fusion technique is used to fuse various emotion identification results that were acquired by LSTM at various lengths of segment-level spectrograms. Thus, the proposed model is very useful to identify the emotional speech of a person.

We did not test the speech of noisy audio data or multilingual speakers' data, or abnormal person speech. Our model may or may not identify the above dataset. But we will focus on multilingual speakers' data to identify the emotion in the future.

#### 5. CONCLUSIONS

This paper presents a novel and highly effective LSTM approach for emotion recognition in audio data. This approach achieves superior performance by combining two key strengths: MFCCs and features extracted from a pre-trained DenseNet model. MFCCs capture essential spectral characteristics of speech, like pitch and intensity, while the DenseNet model, pre-trained on a massive image dataset, learns complex, high-level emotional representations that might not be readily apparent in raw audio. By combining these strengths, our model overcomes the limitations of traditional methods that rely solely on handcrafted features or struggle to capture subtle emotional variations. The proposed approach has been demonstrably successful. It achieves an exceptional accuracy of 100% on the TESS dataset, significantly outperforming models that use only MFCCs or DenseNet features in isolation. This accomplishment signifies a good contribution to the field of affective computing, offering a robust and reliable deep learning model for emotion recognition in audio data. Additionally, delving into methods to improve model interpretability is crucial. For instance, in mental health applications, interpreting the model's decision-making process could provide valuable insights into a user's emotional state. Overall, this work paves the way for the development of more sophisticated emotion recognition



systems that can benefit various fields, covering areas like how people interact with computers, helper programs that understand us, and tools to track our emotional state. By continuing to refine and expand upon this approach, we can create systems that can accurately recognize and develop the ability to perceive and respond to human emotions, paving the way for emotionally intelligent machines that can interact more effectively with users.

## REFERENCES

- [1] Chen, L., Wang, K., Li, M., Wu, M., Pedrycz, W., Hirota, K. (2022). K-means clustering-based kernel canonical correlation analysis for multimodal emotion recognition in human-robot interaction. *IEEE Transactions on Industrial Electronics*, 70(1): 1016-1024. <https://doi.org/10.1109/TIE.2022.3150097>
- [2] Yi, L., Mak, M.W. (2020). Improving speech emotion recognition with adversarial data augmentation network. *IEEE Transactions on Neural Networks and Learning Systems*, 33(1): 172-184. <https://doi.org/10.1109/TNNLS.2020.3027600>
- [3] Atmaja, B.T., Sasou, A. (2022). Evaluating self-supervised speech representations for speech emotion recognition. *IEEE Access*, 10: 124396-124407. <https://doi.org/10.1109/ACCESS.2022.3225198>
- [4] Zhang, L.M., Ng, G.W., Leau, Y.B., Yan, H. (2023). A parallel-model speech emotion recognition network based on feature clustering. *IEEE Access*, 11: 71224-71234. <https://doi.org/10.1109/ACCESS.2023.3294274>
- [5] Lei, Y., Cao, H. (2023). Audio-visual emotion recognition with preference learning based on intended and multi-modal perceived labels. *IEEE Transactions on Affective Computing*, 14(4): 2954-2969. <https://doi.org/10.1109/TAFFC.2023.3234777>
- [6] Jiang, P., Xu, X., Tao, H., Zhao, L., Zou, C. (2021). Convolutional-recurrent neural networks with multiple attention mechanisms for speech emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(4): 1564-1573. <https://doi.org/10.1109/TCDS.2021.3123979>
- [7] Tian, J., She, Y. (2022). A visual-audio-based emotion recognition system integrating dimensional analysis. *IEEE Transactions on Computational Social Systems*, 10(6): 3273-3282. <https://doi.org/10.1109/TCSS.2022.3200060>
- [8] Bhuyan, H.K., Ravi, V., Brahma, B., Kamila, N.K. (2022). Disease analysis using machine learning approaches in healthcare system. *Health and Technology*, 12(5): 987-1005. <https://doi.org/10.1007/s12553-022-00687-2>
- [9] Bhuyan, H.K., Ravi, V. (2023). An integrated framework with deep learning for segmentation and classification of cancer disease. *International Journal on Artificial Intelligence Tools*, 32(2): 2340002. <https://doi.org/10.1142/S021821302340002X>
- [10] Bhuyan, H.K., Vijayaraj, A., Ravi, V. (2023). Diagnosis system for cancer disease using a single setting approach. *Multimedia Tools and Applications*, 82(30): 46241-46267. <https://doi.org/10.1007/s11042-023-15478-8>
- [11] Bhuyan, H.K., Vijayaraj, A., Ravi, V. (2023). Development of secrete images in image transferring system. *Multimedia Tools and Applications*, 82(5): 7529-7552. <https://doi.org/10.1007/s11042-022-13677-3>
- [12] Mohanty, R., Bhuyan, H.K., Pani, S.K., Ravi, V., Krichen, M. (2023). Bird species recognition using spiking neural network along with distance based fuzzy co-clustering. *International Journal of Speech Technology*, 26(3): 681-694. <https://doi.org/10.1007/s10772-023-10040-1>
- [13] Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., Schuller, B.W. (2023). Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 10745-10759. <https://doi.org/10.1109/TPAMI.2023.3263585>
- [14] Shukla, A., Petridis, S., Pantic, M. (2021). Does visual self-supervision improve learning of speech representations for emotion recognition? *IEEE Transactions on Affective Computing*, 14(1): 406-420. <https://doi.org/10.1109/TAFFC.2021.3062406>
- [15] Su, B.H., Lee, C.C. (2022). Unsupervised cross-corpus speech emotion recognition using a multi-source cycle-GAN. *IEEE Transactions on Affective Computing*, 14(3): 1991-2004. <https://doi.org/10.1109/TAFFC.2022.3146325>
- [16] Leem, S.G., Fulford, D., Onnela, J.P., Gard, D., Busso, C. (2023). Selective acoustic feature enhancement for speech emotion recognition with noisy speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32: 917-929. <https://doi.org/10.1109/TASLP.2023.3340603>
- [17] Liu, Z., Kang, X., Ren, F. (2023). Dual-TBNet: Improving the robustness of speech features via dual-transformer-BILSTM for speech emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 2193-2203. <https://doi.org/10.1109/TASLP.2023.3282092>
- [18] Khurana, Y., Gupta, S., Sathiyaraj, R., Raja, S.P. (2022). RobinNet: A multimodal speech emotion recognition system with speaker recognition for social interactions. *IEEE Transactions on Computational Social Systems*, 11(1): 478-487. <https://doi.org/10.1109/TCSS.2022.3228649>
- [19] Sahu, S., Gupta, R., Espy-Wilson, C. (2020). Modeling feature representations for affective speech using generative adversarial networks. *IEEE Transactions on Affective Computing*, 13(2): 1098-1110. <https://doi.org/10.1109/TAFFC.2020.2998118>
- [20] Zhang, S., Zhang, S., Huang, T., Gao, W., Tian, Q. (2017). Learning affective features with a hybrid deep model for audio-visual emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10): 3030-3043. <https://doi.org/10.1109/TCSVT.2017.2719043>
- [21] Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8): 1735-1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- [22] Zhang, S., Zhao, X., Tian, Q. (2019). Spontaneous speech emotion recognition using multiscale deep convolutional LSTM. *IEEE Transactions on Affective Computing*, 13(2): 680-688. <https://doi.org/10.1109/TAFFC.2019.2947464>
- [23] Nasim, A.S., Chowdory, R.H., Dey, A., Das, A. (2021). Recognizing speech emotion based on acoustic features using machine learning. In 2021 International



- Conference on Advanced Computer Science and Information Systems (ICACSIS), Depok, Indonesia, pp. 1-7.  
<https://doi.org/10.1109/ICACSIS53237.2021.9631319>
- [24] Zhou, Y., Liang, X., Gu, Y., Yin, Y., Yao, L. (2022). Multi-classifier interactive learning for ambiguous speech emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 695-705. <https://doi.org/10.1109/TASLP.2022.3145287>
- [25] Turchet, L., Pauwels, J. (2021). Music emotion recognition: intention of composers-performers versus perception of musicians, non-musicians, and listening machines. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 305-316. <https://doi.org/10.1109/TASLP.2021.3138709>
- [26] Gavali, M.P., Verma, A. (2023). Automatic recognition of emotions in speech with large self-supervised learning transformer models. In *2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings)*, Mount Pleasant, MI, USA, pp. 1-7. <https://doi.org/10.1109/AIBThings58340.2023.10292462>