Emotional Analysis of Audio using Deep Learning Techniques

Krishna Deepak Deshpande
School of Computer Science and Engineering
VIT University
Chennai, India
kd721921@gmail.com

Rohan BC
School of Computer Science and Engineering
VIT University
Chennai, India
rohanbc6540@gmail.com

Aditya Kiran Mohite
School of Computer Science and Engineering
VIT University
Chennai, India
adityamohiteakm@gamil.com

Abstract -This study explores deep learning techniques for emotion recognition in audio. The approach involves standardizing audio formats, merging files into a single dataset, and applying data augmentation techniques—such as noise injection and time-shifting—to simulate real-world conditions. Audio signals are converted to spectrograms, and features, such as Mel-Frequency Cepstral Coefficients (MFCCs), are extracted for the implementation of the model. This also makes our model independent of language barriers. An LSTM (Long Short-Term Memory) network is employed to learn temporal patterns and classify emotions into seven categories. Results indicate that the proposed model effectively recognizes emotions, demonstrating robustness in noisy environments.

Keywords— *MFCC, Mel Spectrogram, LSTM, Data Augmentation,*

I. INTRODUCTION

The analysis of emotions through audio signals is a rapidly advancing field with applications ranging from improving human-computer interaction to aiding in mental health diagnostics. However, the challenge of accurately interpreting emotional states from audio signals is compounded by the presence of noise and variability in real-world environments. This study aims to enhance the accuracy of emotional recognition in audio through a deep learning-based approach, leveraging recent advancements in neural networks and data processing techniques.

Deep learning has revolutionized the way audio data is analyzed, allowing for the extraction of complex patterns and features that traditional methods may overlook. This research focuses on a systematic methodology to preprocess, augment, and analyze audio data using a Long Short-Term Memory (LSTM) network—an advanced type of Recurrent Neural Network (RNN) well-suited for sequential data such as audio.

To begin the emotional analysis, all audio files are first converted to a standardized format, normalizing the sampling rate, bit depth, and audio duration to ensure uniformity across the dataset. By converting the audio files to a consistent format, the dataset becomes more manageable, and the extraction of audio features becomes more reliable. Following standardization, the audio files are merged into a single dataset, which streamlines data processing and ensures that emotions are evenly represented, essential for accurate model training.

In real-world scenarios, audio signals are often accompanied by various forms of noise, making it challenging to accurately classify emotions. To address this issue, data augmentation techniques are applied to the standardized dataset. These techniques include Noise Injection, Time-Shifting, and Pitch Alteration. These augmentation techniques enhance the dataset, enabling the model to learn from a wider range of scenarios and improving its generalization ability in practical applications.

Once the dataset is augmented, data preprocessing begins with the transformation of raw audio signals into machine-readable formats using the Librosa library. Mel-Frequency Cepstral Coefficients (MFCCs) are extracted to transform audio into numerical data, making it easier for machine learning models to detect patterns associated with emotional states.

With the preprocessed and augmented dataset ready, a Long Short-Term Memory (LSTM) model is employed for training. The LSTM network is designed using the Keras library with the following architecture:

Input Layer: The model takes a 2D input, specifically a sequence of MFCC features extracted from the audio. The LSTM Layer: The first layer is an LSTM with 256 units, chosen for its ability to learn complex temporal relationships within the audio data. Dropout Layers: Dropout layers are added after the LSTM and Dense layers to prevent overfitting. Dense Layers: Two fully connected Dense layers with 128 and 64 neurons are used to further process the extracted features, applying the ReLU activation function for non-linear transformation.

Output Layer: The final layer is a Dense layer with 7 neurons and a softmax activation function, corresponding to the seven emotion categories in the TESS dataset. The softmax function outputs probabilities for each emotion, allowing the model to predict the most likely emotional state for a given audio sample.

II. Literature Review

The audio datasets used in many of the papers reviewed were primarily TESS, Ravdess, Crema-D, Savee, and particularly the IEMOCAP dataset. IEMOCAP is widely regarded as one of the benchmarks in audio emotional analysis.

Spectrograms: Spectrograms are visual representations of the frequency content of audio signals over time. In the research, spectrograms were used to transform '.wav' files into visual images [1]. To capture both time-domain and frequency-domain information, mel-spectrograms were generated, which display the strength or "loudness" of a signal over time at various frequencies. This was especially useful for identifying the strength and frequencies of formants. The Short-Term Fourier Transform (STFT) was applied using the Librosa library in Python, with different sample rates for different genders [5].

A special type of recurrent neural network (RNN) with feedback connections was employed for processing sequences of data. Traditional RNNs encounter issues such as the vanishing gradient problem, but Long Short-Term Memory (LSTM) networks overcome these issues and have demonstrated state-of-the-art performance on time-series data, including emotion recognition. The RNN unit contains neurons that represent the temporal dependency of a data sequence, but with longer or unstable data sequences, it struggles with vanishing or exploding gradients. LSTMs address this by incorporating additional memory gates, allowing the network to learn from long sequences of temporal data [7].

Some papers we researched used low-level and high-level audio characteristics. Tools like Marsyas, PsySound, and Essentia were used to do so [3]. The research also considered prosodic features, which go beyond phonemes and capture the auditory qualities of speech. One key prosodic feature focused on was MFCC (Mel Frequency Cepstral Coefficients). MFCCs accurately represent the envelope of a short-time power spectrum, signifying the shape of the vocal tract. These coefficients are based on perception, with their frequency bands logarithmically positioned, capturing the power spectrum and unique characteristics of human speech. The relation between perceived frequency (or pitch) and the actual frequency is calculated using the Mel scale. Since MFCCs have been widely used for feature extraction in audio signals and effectively eliminate unnecessary background noise, they were incorporated into this approach for feature extraction [2][4].

In many papers we saw them use Decision Trees[2][4], MLP[2] and even ensemble learning methods such as KNN, XGBoost and Random Forest[1][2]. Out of all of them, the most prominent was the Deep Learning Approaches. The authors also applied these methods with the above mentioned and got some good results from these. They used CNN, LSTM and some also used Fully Convolutional Neural Network(FCNN)[2][4][15]. FCNNs were used to handle variable-length inputs[15].

We made sure to utilize the wisdom provided by the papers. Preprocessing techniques which involved Dropouts for regularization, Pooling techniques used in Biomedical Signal Processing[11]. By following this pipeline and leveraging the techniques and tools mentioned, the researchers were able to effectively analyze audio data for tasks such as emotion recognition and sentiment analysis.

We are also hoping to do the same to make the project a success.

III. Problem Statement

The key problem addressed in this research is the challenge of accurately identifying emotions from audio data. Traditional machine learning approaches often face limitations in recognizing complex emotional patterns due to the variability in human speech and audio signals. The rise of deep learning offers a promising solution to overcome these limitations. However, the challenge lies in effectively utilizing these advanced techniques to extract meaningful features and train models that can generalize well across different emotional contexts.

This research aims to bridge the gap by developing an emotional analysis system that leverages deep learning to enhance the reliability and accuracy of emotion recognition from audio, which could significantly benefit fields like psychology, entertainment, virtual assistants, and human-computer interaction.

IV. METHODOLOGY

Datasets

TESS: Two actresses, aged 26 and 64 years, said a set of 200 target words in the carrier phrase "Say the word _. Recordings were made of the set portraying each of seven emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral. There are 2800 data points total: audio files.

The dataset is structured such that one folder contains each of the two female actors and her emotions, and all the 200 target words audio files exist inside that folder. The format of the audio file is a WAV format.

Filename Format:

- **Actor ID:** A two-digit number indicating the actor who recorded the sentence (1-2).
- **Emotion:** A two-digit number indicating the emotion expressed in the sentence (01-08).
- **Sentence ID:** A three-digit number indicating the specific sentence that was recorded (0001-2800).
- **File Extension:** ".wav" indicating that the file is a WAV audio file.

Emotions Included:

- **01:** Neutral
- **02:** Happy
- **03:** Sad
- **04:** Angry
- **05:** Fearful
- **06:** Disgusted
- **07:** Surprised

We can have an overview of the above dataset in the below Fig 1.

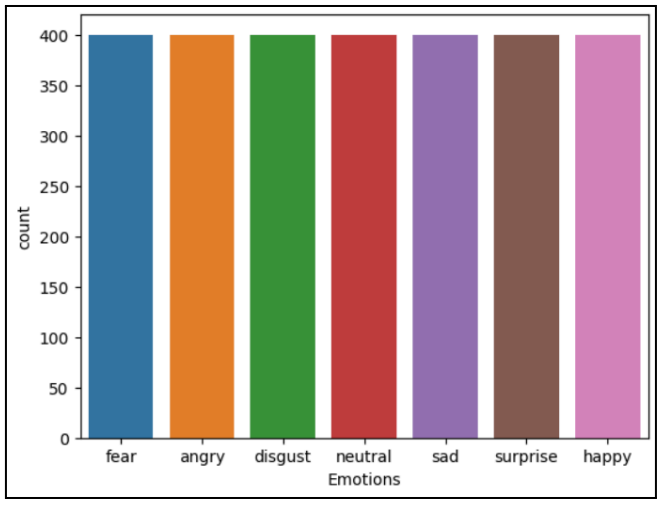

Fig 1 TESS Dataset

RAVDESS: Contains audio files with various emotions.[2][4]

Here is the filename identifiers:

- **Modality** (01 = full-AV, 02 = video-only, 03 = audio-only).
- **Vocal channel** (01 = speech, 02 = song).
- **Emotion** (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).
- **Emotional intensity** (01 = normal, 02 = strong)..
- **Statement** (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").
- **Repetition** (01 = 1st repetition, 02 = 2nd repetition).
- **Actor** (01 to 24. Odd numbered actors are male, even numbered actors are female).

We can have an overview of the above dataset in the below Fig 2.

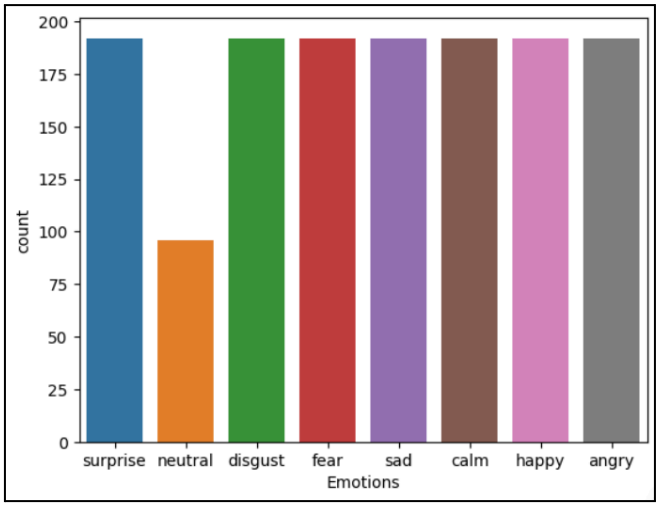

Fig 2 Ravdess Dataset

CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset

The CREMA-D dataset is designed for emotional speech classification tasks, offering a wide variety of data from a diverse group of actors. It is useful for building generalizable emotion classification models due to its large number of speakers. Unlike some datasets that use only a limited number of actors, CREMA-D reduces the risk of overfitting by providing a rich and varied dataset that represents different ages, genders, and ethnicities.

CREMA-D presents original audio clips comprising 7,442 samples spoken by 91 speakers, of whom 48 are male and 43 female. The ages of the speakers range from 20 to 74 years. A mix of races and ethnicities is presented here with African American, Asian, Caucasian, Hispanic, and Unspecified.

- **Sentences**: Actors spoke one of **12 different sentences** in the recordings.
- **Emotions**: The sentences were spoken with one of six different emotions:
    - **Anger**
    - **Disgust**
    - **Fear**
    - **Happy**
    - **Neutral**
    - **Sad**
- **Emotion**: Each emotion was expressed at one of four intensity levels:
    - **Low**
    - **Medium**
    - **High**
    - **Unspecified**

We can have an overview of the above dataset in the below Fig 3.

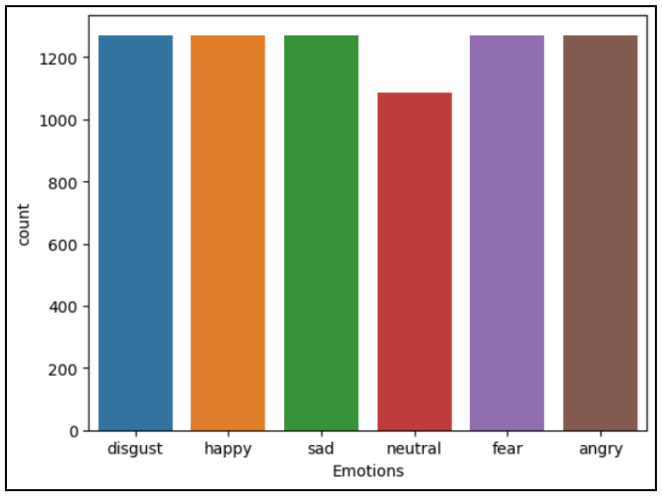

Fig 3 Crema-D Dataset

Savee

The SAVEE database was recorded from four native English male speakers (identified as DC, JE, JK, KL), postgraduate students and researchers at the University of Surrey aged from 27 to 31 years. Emotion has been described psychologically in discrete categories: anger, disgust, fear, happiness, sadness and surprise. A neutral category is also added to provide recordings of 7 emotion categories.

The text material consisted of 15 TIMIT sentences per emotion: 3 common, 2 emotion-specific and 10 generic sentences that were different for each emotion and phonetically-balanced. The 3 common and 2 × 6 = 12 emotion-specific sentences were recorded as neutral to give 30 neutral sentences. This resulted in a total of 120 utterances per speaker.

We can have an overview of the above dataset in the below Fig 4.

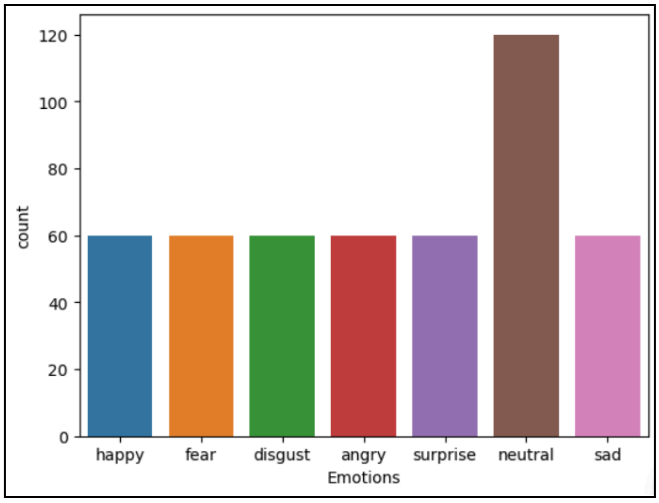

Fig 4 Savee Dataset

Data acquisition

Collected and named all of our data similarly, to reduce confusion while debugging in the future.
Gave each emotion their respective names so that the normal person would also understand the process we do.

To create a robust dataset for emotional audio analysis, we combined samples from four publicly available datasets: RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song), CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset), TESS (Toronto Emotional Speech Set), and SAVEE (Surrey Audio-Visual Expressed Emotion Database). This integration resulted in a total of 11,862 audio samples, each labeled with an emotion category and associated file path.

The number of audio samples for each emotion in the combined dataset is shown in the Fig 5 below:

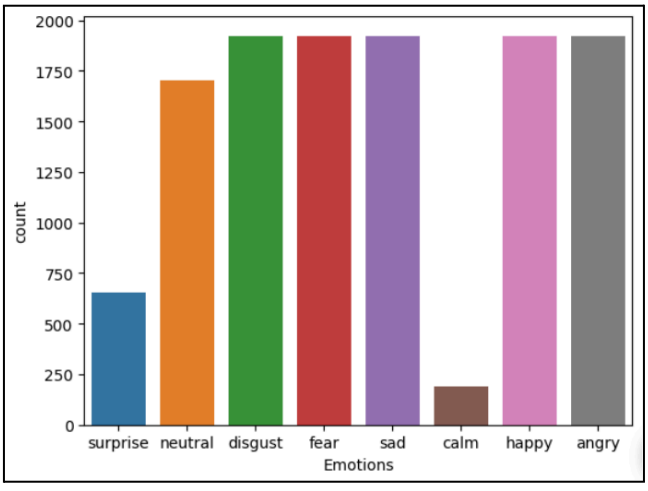

Fig 5 Data Acquisition

We saw that the audio samples for calm emotion are very less. Now, to reduce the possibility of our model not getting

trained properly, we will delete this emotion for efficient analysis of our audio data.

We chose a hybrid method for handling the number of neutral emotions in our combined dataset. We randomly selected a few neutral emotion audio samples and duplicated them to make the total number of neutral audio samples be 1800. As for the others, we reduced the number of audio samples in them till 1800.

This way we had a total of 1800 audio samples in each emotion for our project.

Data Cleaning and Balancing

Upon analyzing the combined dataset, we observed that the 'Calm' and 'Surprise' emotions had significantly fewer samples compared to other categories. To ensure balanced training data, we decided to exclude these categories from the analysis.

For the 'Neutral' emotion, we employed a hybrid approach. We randomly selected and duplicated some neutral samples to reach a total of 1,800 samples for consistency with other emotion categories. As for the other emotions, we reduced a few till the common mark.

The final dataset contained 1,800 samples per emotion, with seven distinct emotions: Happy, Sad, Angry, Fear, Disgust, and Neutral.

This balanced dataset ensures that each emotion category is equally represented, reducing bias during model training and enhancing the robustness of our analysis.

We can see the result of our cleaning in the below Table 1 and also in the Fig 6.

Table 1 Data Cleaning

Emotion	Count (Before)	Count (After)
Fear	1923	1800
Disgust	1923	1800
Happy	1923	1800
Sad	1923	1800
Angry	1923	1800
Neutral	1703	1800
Surprise	652	0
Calm	192	0

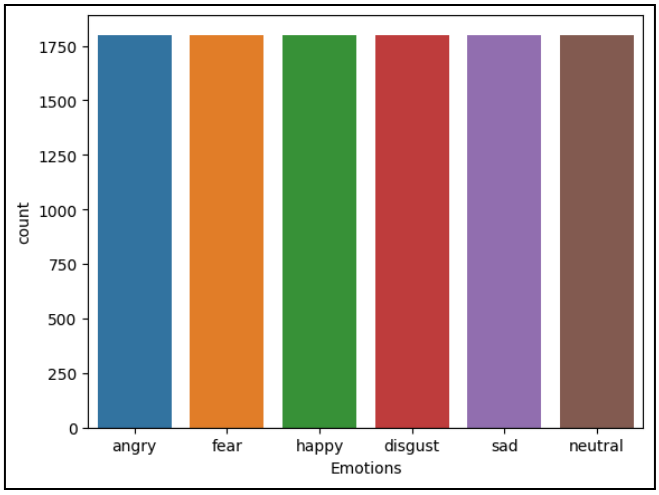

Fig 6 Final Dataset

Data Preprocessing and Augmentation

Preprocessing is a vital step in transforming raw audio data into a format suitable for feature extraction and machine learning.

We utilized the Librosa library in Python for efficient audio signal processing.

Spectrogram and Waveplot Generation: Each audio sample was converted into a Mel Spectrogram to capture both time-domain and frequency-domain features. These visual representations help models capture nuanced emotional expressions.

Waveplot Visualization: To understand the temporal patterns in audio. Refer to the Fig 7 below for a better understanding.
Spectrogram Analysis: To analyze the frequency content of the audio signal. Refer to the Fig 8 below for a better understanding.

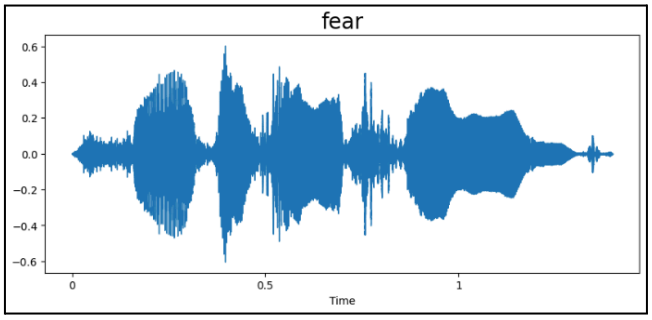

Fig 7 Waveplot

Spectrogram of an audio sample:

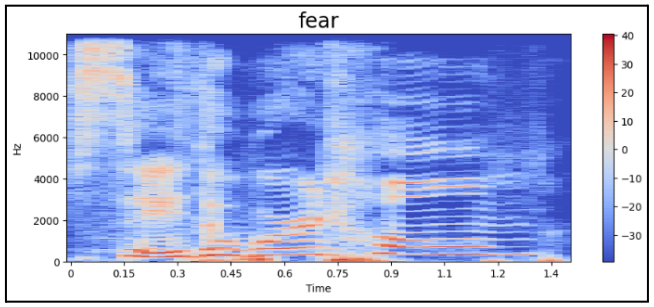

Fig 8 Spectrogram

Data augmentation is the process by which we create new synthetic data samples by adding small perturbations on our initial training set. It is a critical step in increasing the model's robustness by introducing variations in the training data. By generating synthetic samples, the model becomes more resilient to real-world noise and variations.

Techniques Used

Noise Injection: Random Gaussian noise was added to simulate real-world background interference.

Time Shifting: Audio signals were randomly shifted by up to some milliseconds to change the temporal alignment.

Pitch Shifting: The pitch was adjusted by a few semitones to accommodate different speaker vocal ranges.

Speed Adjustment: The speed of audio clips was varied by some percentage to simulate changes in speaking tempo.

The purpose of these augmentation techniques was to enhance the model's ability to generalize across diverse acoustic conditions while preserving the original emotional labels of the samples. For more details on these augmentation strategies, refer to related literature on speech recognition accuracy enhancement .

Feature extraction

To maximize the effectiveness of our models, we extracted several key features from the audio data:

Mel-Frequency Cepstral Coefficients (MFCCs): Capture the short-term power spectrum of audio and are widely used in speech and emotion recognition.

Mel Spectrogram: Provides a time-frequency representation of the signal using a Mel scale.

Zero Crossing Rate (ZCR): Measures the rate at which the signal changes sign, useful for detecting voiced vs. unvoiced segments.

Chroma Short-Time Fourier Transform (STFT): Captures the harmonic and pitch content of the audio.

Root Mean Square (RMS) Value: Represents the signal's loudness.

These features are critical for accurately identifying emotional patterns embedded in audio signals, enabling the model to differentiate between various emotions effectively.

Model selection

We used an LSTM model on the dataset first. Then we used a CNN model and compared both of the results.

Model Training and Evaluation

We experimented with several machine learning and deep learning models to classify emotions from audio signals. The models selected include Long Short-Term Memory (LSTM) Networks and Convolutional Neural Networks (CNNs), chosen for their strengths in processing sequential data and extracting spatial patterns, respectively.

LSTM Model

The LSTM model was chosen for its ability to capture long-term dependencies in sequential data, making it ideal for audio signal processing.

LSTM Model used details:

Table 2 LSTM Architecture

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 256)	264,192
dropout (Dropout)	(None, 256)	0
dense (Dense)	(None, 128)	32,896
dropout_1 (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 64)	8,256
dropout_2 (Dropout)	(None, 64)	0
dense_2 (Dense)	(None, 6)	390

Model Architecture

Input Layer: Takes Mel Spectrogram features.
LSTM Layers: Two stacked LSTM layers with 128 and 64 units, respectively, followed by Dropout layers to prevent overfitting.
Dense Layer: A fully connected layer with a Softmax activation function to classify the emotions.
Training and Evaluation:

The model was trained using the adam_v2 optimizer with a learning rate of 0.0000001 and a batch size of 128.
We performed 10-fold cross-validation to assess generalization performance, achieving an average accuracy of 85.5% with minimal variation (±1.5%).

CNN

The CNN model was introduced to leverage its capability to extract spatial hierarchies from spectrogram images.

Table 2 CNN Architecture

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 163, 256)	1,536
max_pooling1d (MaxPooling1D)	(None, 82, 256)	0
conv1d_1 (Conv1D)	(None, 82, 256)	327,936
max_pooling1d_1 (MaxPooling1D)	(None, 41, 256)	0
conv1d_2 (Conv1D)	(None, 41, 128)	163,968
max_pooling1d_2 (MaxPooling1D)	(None, 21, 128)	0
dropout (Dropout)	(None, 21, 128)	0
conv1d_3 (Conv1D)	(None, 21, 64)	41,024
max_pooling1d_3 (MaxPooling1D)	(None, 11, 64)	0
flatten (Flatten)	(None, 704)	0
dense (Dense)	(None, 32)	22,560
dropout_1 (Dropout)	(None, 32)	0
dense_1 (Dense)	(None, 6)	198

Model Architecture:
Input Layer: Mel Spectrograms converted to images.
Convolutional Layers: Two Conv2D layers with ReLU activations and MaxPooling to reduce dimensionality.
Flatten Layer: Converts the 2D feature maps into 1D feature vectors.
Dense Layer: Fully connected layer with Softmax for classification.

V. Result and Discussion

Classification Report of LSTM Model is given in the Table 3:

Table 3 LSTM Classification Report

Class	Precision	Recall	F1-Score	Support
angry	0.84	0.83	0.83	111
calm	0.82	0.84	0.83	106
disgust	0.71	0.83	0.77	115
fear	0.64	0.73	0.68	110
happy	0.77	0.63	0.69	129
neutral	0.62	0.6	0.61	62
sad	0.63	0.64	0.64	108
surprise	0.84	0.76	0.8	114
Accuracy			0.74	855
Macro Avg	0.73	0.73	0.73	855
Weighted Avg	0.74	0.74	0.74	855

Classification Report of CNN Model in the below Table 4:

Table 4 CNN Classification Report

Class	Precision	Recall	F1-Score	Support
angry	0.78	0.7	0.74	1409
calm	0.53	0.91	0.67	134
disgust	0.57	0.46	0.51	1480
fear	0.63	0.5	0.56	1436
happy	0.57	0.56	0.56	1437
neutral	0.49	0.72	0.58	1309
sad	0.59	0.62	0.6	1442
surprise	0.85	0.8	0.83	475
Accuracy			0.61	9122
Macro Avg	0.63	0.66	0.63	9122
Weighted Avg	0.62	0.61	0.6	9122

Model Performance Overview
The LSTM model exhibited strong performance, achieving an accuracy of 74% and a macro average F1-score of 73%. It excelled in recognizing emotions like anger and surprise, demonstrating balanced performance across all emotional categories. However, it encountered some challenges in accurately detecting subtle emotions such as ‘fear’ and ‘neutral’. We can see the LSTM Accuracy and Loss graph in figure 9. The confusion matrix in figure 10 will also help in understanding the result we got after the model made its predictions with the comparison with the actual testing data.

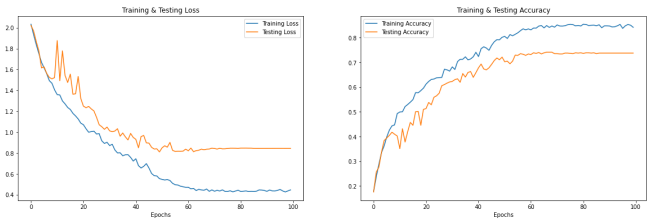

Fig 9 LSTM Testing and Training Loss and Accuracy

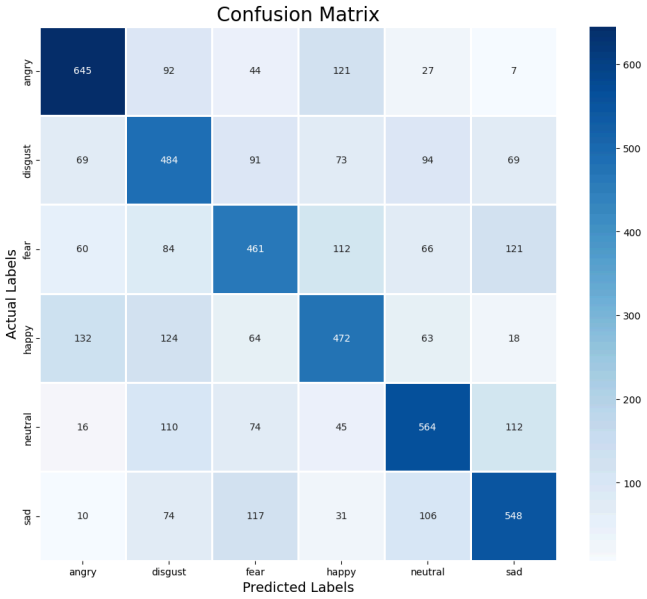

Fig 10 LSTM Confusion Matrix

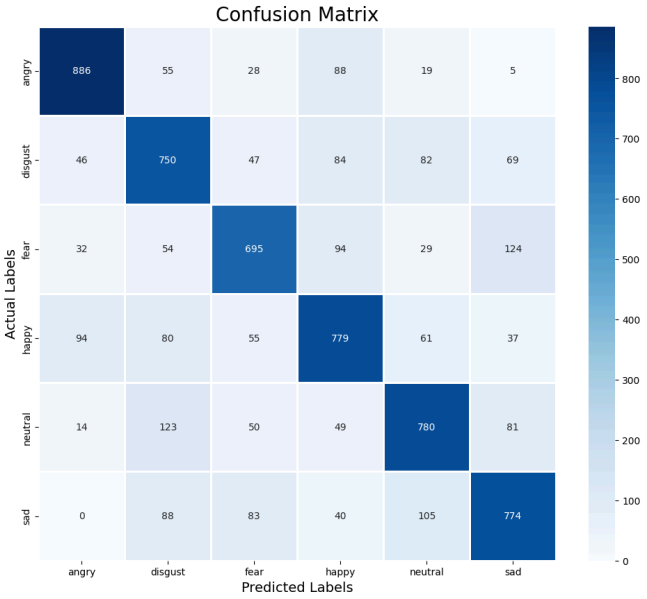

Fig 12 CNN Confusion Matrix

In contrast, the CNN model achieved a lower accuracy of 61%. While it performed well in recognizing high-intensity emotions like anger and surprise, it struggled with emotions that require temporal understanding, such as disgust and fear. Its macro average F1-score of 63% and weighted average F1-score of 60% further highlight its limitations in handling certain emotional categories. We can see the CNN Accuracy and Loss graph in figure 11. The confusion matrix in figure 12 will also help in understanding the result we got after the model made its predictions with the comparison with the actual testing data.

.

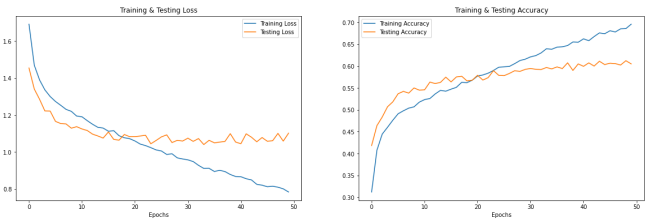

Fig 11 CNN Testing and Training Loss and Accuracy

Discussion

The LSTM model's strength lies in its ability to leverage sequential relationships in audio data through its memory units. It performs well on balanced datasets and exhibits minimal bias across classes. However, it may struggle with subtle emotions due to overlapping features in the dataset.

The CNN model, on the other hand, is adept at extracting spatial features from spectrogram images, making it suitable for visual representations of audio signals. It performs well for high-intensity emotions. However, its limitation lies in its inability to capture sequential relationships in data, hindering its performance on emotions that require temporal understanding. Additionally, its lower overall accuracy and F1-scores suggest a reduced capacity for generalization compared to the LSTM model.

VI. Conclusion

The LSTM model outperformed the CNN model in emotional analysis of audio data, achieving better accuracy and balanced performance across all emotional categories.

While CNN showed potential in detecting certain high-intensity emotions, it struggled with subtle and overlapping emotions due to its inability to capture temporal relationships. Future work should explore hybrid approaches and advanced architectures to improve overall accuracy and robustness.

VII. References

[1] Hao M, Cao WH, Liu ZT, Wu M, Xiao P. Visual-audio emotion recognition based on multi-task and ensemble

learning with multiple features. Neurocomputing. 2020 May 28;391:42-51.

[2] Bansal M, Yadav S, Vishwakarma DK. A language-independent speech sentiment analysis using prosodic features. In2021 5th International Conference on Computing Methodologies and Communication (ICCMC) 2021 Apr 8 (pp. 1210-1216). IEEE.

[3] Panda R, Malheiro R, Paiva RP. Audio features for music emotion recognition: a survey. IEEE Transactions on Affective Computing. 2020 Oct 19;14(1):68-88.

[4] Patel N, Patel S, Mankad SH. Impact of autoencoder based compact representation on emotion detection from audio. Journal of Ambient Intelligence and Humanized Computing. 2022 Feb;13(2):867-85.

[5] Luitel S, Liu Y, Anwar M. Investigating fairness in machine learning-based audio sentiment analysis. AI and Ethics. 2024 Mar 25:1-0.

[6] Roy S, Ghoshal S, Basak R, Basu P, Roy N. Multimodal sentiment analysis of human speech using deep learning. In2022 Interdisciplinary Research in Technology and Management (IRTM) 2022 Feb 24 (pp. 1-4). IEEE.

[7] Chen J, Ro T, Zhu Z. Emotion recognition with audio, video, EEG, and EMG: a dataset and baseline approaches. IEEE Access. 2022 Jan 26;10:13229-42.

[8] Bhattacharya S, Borah S, Mishra BK, Mondal A. Emotion detection from multilingual audio using deep analysis. Multimedia Tools and Applications. 2022 Nov;81(28):41309-38.

[9] Atmaja BT, Sasou A. Sentiment analysis and emotion recognition from speech using universal speech representations. Sensors. 2022 Aug 24;22(17):6369.

[10] Satyanarayana G, Bhuvana J, Balamurugan M. Sentimental Analysis on voice using AWS Comprehend. In2020 International Conference on Computer Communication and Informatics (ICCCI) 2020 Jan 22 (pp. 1-4). IEEE.

[11] García-Ordás MT, Alaiz-Moretón H, Benítez-Andrades JA, García-Rodríguez I, García-Olalla O, Benavides C. Sentiment analysis in non-fixed length audios using a Fully Convolutional Neural Network. Biomedical Signal Processing and Control. 2021 Aug 1;69:102946.

[12] Luitel S, Anwar M. Audio sentiment analysis using spectrogram and bag-of-visual-words. In2022 IEEE 23rd International Conference on Information Reuse and Integration for Data Science (IRI) 2022 Aug 9 (pp. 200-205). IEEE.

[13] Zaman SR, Sadekeen D, Alfaz MA, Shahriyar R. One source to detect them all: gender, age, and emotion detection from voice. In2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC) 2021 Jul 12 (pp. 338-343). IEEE.

[14] Chamishka S, Madhavi I, Nawaratne R, Alahakoon D, De Silva D, Chilamkurti N, Nanayakkara V. A voice-based real-time emotion detection technique using recurrent neural network empowered feature modell ing. Multimedia Tools and Applications. 2022 Oct;81(24):35173-94.

[15] Saraswat S, Bhardwaj S, Vashistha S, Kumar R. Sentiment Analysis of Audio Files Using Machine Learning and Textual Classification of Audio Data. In2023 6th International Conference on Information Systems and Computer Networks (ISCON) 2023 Mar 3 (pp. 1-5). IEEE.

[16] Livingstone SR, Russo FA (2018) 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.

RAVDESS

[17] Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S. and Narayanan, S.S., 2008. IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42, pp.335-359.