Enhancing Human Emotion Detection in Audio Data with Deep Neural Networks Using Cross-Dataset

Sandushi Weraduwa
Undergraduate of Department of Industrial management
University of Kelaniya
Sri Lanka
sandushiw98@gmail.com

Minuli Kannangara
Undergraduate of Department of Industrial management
University of Kelaniya
Sri Lanka
kkminuli@gmail.com

Shiran SuriyaPathiraja
Undergraduate of Department of Industrial management
University of Kelaniya
Sri Lanka
shiranay22083@gmail.com

***Abstract*—Understanding emotions from spoken language is a natural human ability, but it brings machines into a great perplexity due to the complex variation in an individual's speech. Our research focuses on working out this challenge by developing a model that makes use of deep neural networks, particularly Convolutional Neural Networks(CNN), for the recognition of emotions from speech. The goal is to be equipped with a system capable of independent emotion detection for a variety of audio datasets, adaptation to new situations, and staying accurate even in fully new cases of data.To achieve this, we combined two strong datasets: the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and the Toronto Emotional Speech Set (TESS). The TESS dataset is particularly useful because it includes high-quality recordings from female speakers, helping to balance the gender differences often found in other datasets. This balance makes our model better at recognizing emotions across different voices. We implemented state-of-the-art techniques in audio feature extraction so that the input to our model only consists of relevant features, enabling it to extract the emotional content correctly. Our model was trained on eight emotions: neutral, calm, happy, sad, angry, fearful, disgusted, and surprised. For measuring model performance, we used the F1 score and obtained a weighted average of 0.94 on the test set. It did best on classifying the "Calm" emotion, with an accuracy of 0.97, and worst for the "Sad" emotion, with an accuracy of 0.91. Notwithstanding, our model outperforms current models in handling the huge diversity of voices in speech data.**

***Index Terms*—Classification, CNN, Deep Learning, Machine Learning, Mel-Frequency Cepstral Coefficients, RAVDESS, Speech Emotion Detection, TESS**

I. Introduction

A. Background of the Problem

Since the earliest days of human society, Verbal communication has been playing a significant role in the enhancement of societies. But even before we started talking, emotions were there staging themselves as one of the first nondiscursive means of communication with no words to say. It was also possible to demonstrate how a person felt or what he/she was thinking using body language. To date, emotions are still relevant in the conversation because they enrich it and make it clearer. Often, we notice feelings through facial movement or the pitch and amplitude of a human voice but something challenging for a machine. As it now stands, computers can translate the spoken word into writing better than before, but they are far from getting the tone of the conversation. Modeling emotions within speech is one of the greatest barriers to the conversation between man and machine. That is why our work is devoted to finding the best strategies for identifying the primary emotional states of a person when he speaks.

B. Purpose of the Study

The main purpose of this study is to explore and develop a robust model for emotion recognition from spoken language, focusing exclusively on audio data. In the literature, it is shown how individuals often express more than one basic emotion simultaneously [1] (six Ekman's emotions [2]). In this work, we focus on the problem of determining which of the emotions is the most dominant in an audio track, rather than attempting to recognize different combinations of emotions. This makes the task of recognizing emotions relatively easier in terms of complexity and more feasible for implementation in human-machine interaction systems. The study employs Mel-frequency cepstral coefficients (MFCC) as the primary feature extraction technique [3], a method well-suited for analyzing audio signals in emotion classification tasks. The objective of this work is to improve the accuracy and reliability of emotion detection models based purely on audio data so that their contribution could be extended to support, more generally, emotion-aware human-computer interaction.

C. Objectives

The primary objectives of this research are as follows:

1) To develop a Versatile Emotion Classification Model: Develop a system that involves the use of a neural-based network in the detection of emotions from various audio datasets and ensure the data set is flexible enough to adapt to various characteristics of the data set.
2) To enhance Model Generalization: Emphasis the issue of achieving high accuracy on unseen inputs and reducing fluctuations in the model's performance depending on the type of audio data.
3) To implement Advanced Feature Extraction Techniques: Employ sophisticated signal processing methods to extract robust features from audio data that are effective across diverse datasets and help in accurately detecting emotions.

The objective of this study is to cover all the gaps in existing research by proposing a comprehensive model that not only excels in emotion recognition from audio data but also adapts effectively to diverse and unseen datasets. By focusing on versatility, generalization, and advanced feature extraction, this study aims to push the boundaries of current emotion recognition technologies, offering a more reliable and inclusive solution for human-computer interaction.

II. Literature review

Advances in the deep learning approaches to extract emotions from speech data bring long-term potential to a great variety of areas concerning human-to-computer interactions, diagnosis and therapy of mental disorders, and more. SER has seen a massive improvement in the recent past with the help of new developments in deep learning architectures and feature extraction techniques. In the present literature review, three groundbreaking research articles are discussed, which aim at the improvement of emotion detection via new neural network models and approaches. Specifically, this literature review is intended to analyze the current research state in speech emotion recognition (SER), with a particular focus on how deep learning principles have been fine-tuned in several contexts. To provide a structured and in-depth analysis, the review is divided into several critical sub-topics under the relevant research areas: Emotion recognition using deep learning architectures, extraction and analysis of audio features, the robustness of SER system, effect of dataset characteristics, real-time emotion detection, and effect of demographic factors.

A. *Deep Learning for Emotion Recognition*

Deep learning techniques, when combined with traditional audio feature processing, can significantly enhance SER, demonstrating an improvement of 3% in classification accuracy, reaching up to 92.55% on multi-corpus datasets. However, this approach might be less effective in real-time and multilingual settings due to its computational demands and limited language diversity handling [4]. A combined approach using CNN and bi-directional long short-term memory (Bi-LSTM) has proven effective for emotional state classification from speech, benefiting from processing both spatial and temporal data, leading to superior performance compared to conventional methods. Despite its accuracy, the complexity of the CNN-BiLSTM architecture complicates the interpretability of the model, posing challenges in understanding its internal decision-making processes [5]. Various deep learning models, including CNNs, RNNs, and CNN-RNN hybrids, are explored for their effectiveness in detecting emotions in speech and addressing practical applications such as customer service and mental health, with a reported accuracy of 75%. These models show potential, though they require further refinement to improve the precision of emotion classification [6]. Research into the impact of sample duration and linguistic variations (English vs. Italian) reveals differing levels of emotion detection accuracy, ranging from 62% to 92%, underscoring how audio sample length and language diversity can affect the effectiveness of deep learning models in emotion recognition. Despite promising results, the research is limited to two languages, which may not adequately represent the variability in emotional expression across different cultural contexts [7].

B. *Audio Feature Extraction and Audio Feature Analyzer*

Hence accurate emotion recognition in audio data is highly dependent on feature extraction of powerful audio features including pitch and fractional cepstral coefficients. These features are important to determine several emotions as they offer other useful information related to the pitch and duration of the audio signals I/E, which are pertinent to the execution of emotion models. Further, the audio feature analyzer is also ranked crucial in improving the emotion recognition rate. Because our new approach extracts audio features, it improves algorithms for detection and increases overall system performance. Recent developments in this technology have made it possible for the recognition rate to go as high as 81%. At this rate, it was 81.67 % proving that this model is efficient in real-time emotion detection through sound signals. [8].

C. *Noise robustness*

Research in enhancing the resilience of emotion recognition systems to environmental noise and recording variations has led to innovative approaches aimed at improving noise robustness. One technique explored involves increasing the recognition accuracy of speech impacted by unpredictable noise through the addition of a known, controlled masking noise. This method is particularly effective in scenarios with crosstalk corrupting noise, where it has been shown to achieve a significant improvement in recognition accuracy by over 24% [9]. By integrating the same type of masking noise into both training and test data, the discrepancy between these conditions is minimized, ensuring consistency regardless of the nature or consistency of the corrupting noise. Additionally, another study focuses on tackling the challenge of noise robustness in speech recognition by utilizing the concept of missing frames. This approach proposes a noise-robust, yet computationally efficient system for recognizing speech within a small vocabulary setting [10], demonstrating

significant advancements in handling noise interference in speech recognition tasks.

D. Artificial Neural Networks

When discussing the field of artificial intelligence [11] explores the application of various artificial neural network (ANN) architectures, including recurrent neural networks (RNNs) and convolutional neural networks (CNNs), to identify emotions in speech audio clips. Utilizing a dataset annotated with specific emotional states, the study effectively trains ANNs to achieve significant accuracy metrics evaluated through precision, recall, and confusion matrices. However, the focus on a single emotional dataset and reliance on traditional acoustic features may not adequately capture the full range of emotional variability across different cultures and languages, suggesting a need for more diverse dataset incorporation in future research.

E. Convolutional Neural Network (CNN)

Deep learning systems, particularly CNNs, are leveraged to infer hierarchical representations of input data, enhancing the categorization of emotional content in speech. This paper introduces a novel approach using semi-CNNs designed specifically for SER. The semi-CNN architecture undergoes a two-stage training process, focusing initially on fundamental emotional features before advancing to more complex recognition tasks. This methodology allows for a refined extraction and classification of emotional states from audio data, showcasing the unique capabilities of semi-CNNs in handling nuanced emotional expressions [12]. Another study [13] explores the application of Convolutional Neural Networks (CNNs) to the task of distinguishing genuine from recaptured audio recordings. By incorporating a preliminary feature preprocessing based on Electric Network Frequency (ENF) analysis, the CNN model is optimized for handling small audio clips, typically of 2-second duration, which pose a significant challenge for existing technologies. A performance comparison against the support tensor machine highlights CNN's superior capability in audio recapture detection, demonstrating its effectiveness even in minimal and challenging audio samples.

F. Recurrent Neural Networks (RNNs)

Recurrent Neural Networks (RNNs) are very useful in dealing with the temporal structure of speech and this is desirable when modeling the dynamics of speech in the emotion detection process. These improve the accuracy of the depiction of the interactional nature of speech and, thereby, the system's capacity for capturing the definitive, variations in affect across temporal intervals and are particularly optimal for speech imitation and identification. On the same note, it has been established that a good self-training algorithm for RNN speech models can greatly enhance the model's capacity to extract relevant features from the signals. The outcomes obtained in experiments prove that this new strategy is helpful in the identification of isolated speech words with the help of the phonetic division of speech [14].

G. Real-time Emotion Detection

Developing real-time emotion detection systems is crucial for applications requiring immediate response, such as virtual assistants and customer service bots. Studies have shown that using Bayesian Quadratic Discriminate Classifiers can provide effective low-latency performance in emotion detection, handling the challenge of feature overlap across different emotion classes [15].In fact another study [16]presents a system for real-time speech emotion recognition that employs deep learning models such as MLP, CNN, and a hybrid CNN-BiLSTM, enhanced by data augmentation techniques like noise addition and spectrogram shifting. The system exhibits high accuracy across multiple datasets, with the CNN-BiLSTM model achieving the highest average accuracy rates: 99.90% on TESS, 98.76% on EmoDB, and 90.09% on RAVDESS. Despite these impressive results, the study notes challenges in detecting human emotions during real-time human-machine interactions, potentially limiting its efficacy in dynamic or uncontrolled environments.

H. Impact of Gender on Emotion Detection

Gender is considered as an independent variable and gender differences in the perception and detection of emotions in speech are discussed in detail and analyzed for a better understanding of how they are distorted. This demographic factor also has a significant impact on the making of emotion recognition systems more sensitive and correct by responding to gender differences within people [17].

Critical Gap

Among the research gaps recognizable in the current literature, one of the most important ones is the absence of SER models designed for high work in several datasets. Such current models as the ones trained on RAVDESS, perform well as long as the model is tested on the same dataset but if put into other datasets like TESS, the accuracy sharply drops. They describe an issue of dataset bias in which models have learned to rely on certain features and emotional clues that are inherent in the training set. Thus far, there isn't any available SER model that identifies and solves the aforementioned problem of data selection and combination, in a way that would address the issues related to the usage of RAVDESS or TESS databases, the goal being to increase the generality of the model. This has gaps to fill because of the need to build SER systems that are accurate and also resistant to noise and other variations in the inputs.

III. THE PROPOSED MODEL

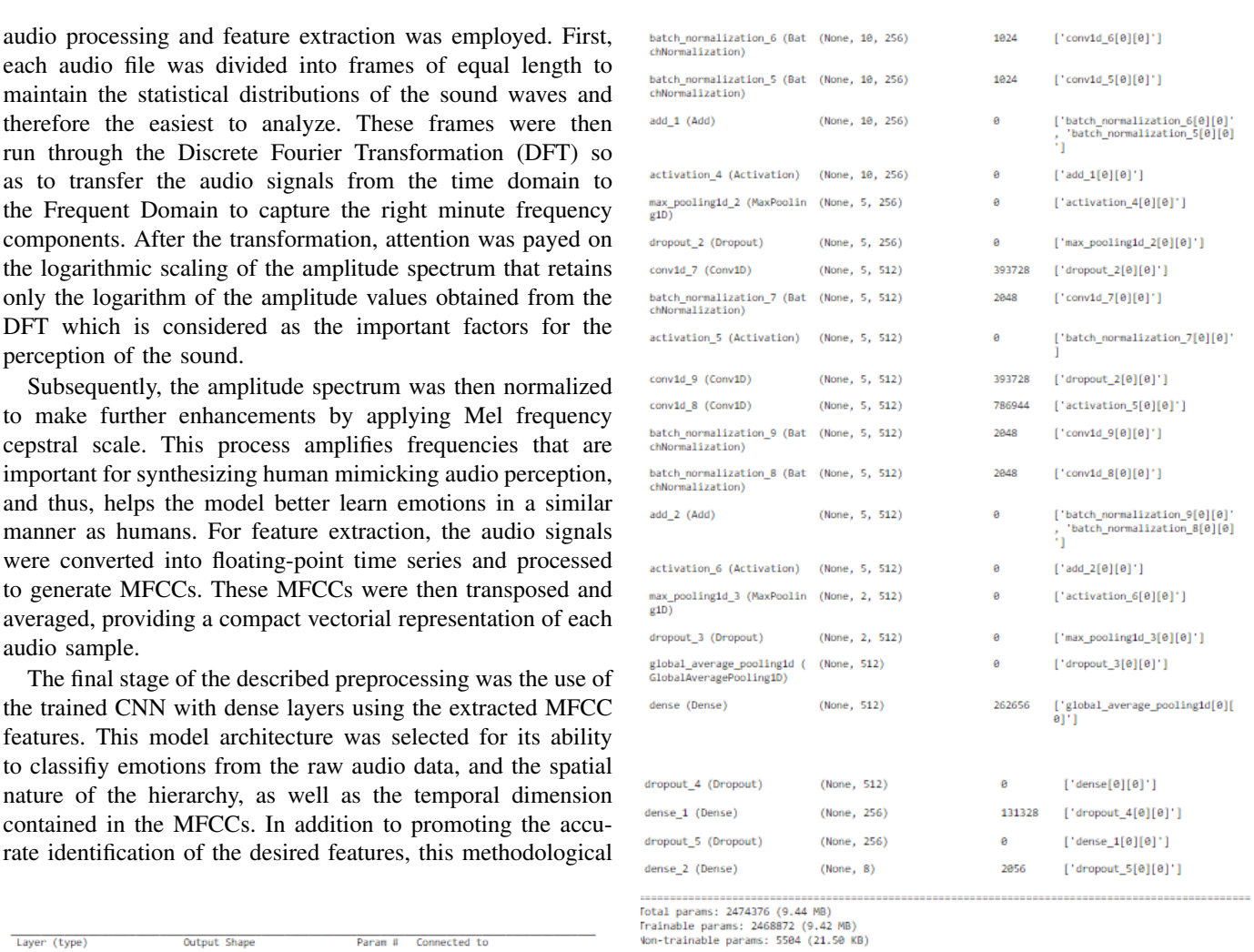
The proposed model focuses mainly on CNN and dense layers. The first aim is to attempt to train this model by using MFCC as the only feature vector. MFCCs represent features extracted from the logarithm of the power spectra density of an audio signal after passing through a Mel-filter bank which simulates the human auditory response system. [18]

In the case of developing the emotion recognition model, a systematic approach involving techniques such as advanced

audio processing and feature extraction was employed. First, each audio file was divided into frames of equal length to maintain the statistical distributions of the sound waves and therefore the easiest to analyze. These frames were then run through the Discrete Fourier Transformation (DFT) so as to transfer the audio signals from the time domain to the Frequent Domain to capture the right minute frequency components. After the transformation, attention was payed on the logarithmic scaling of the amplitude spectrum that retains only the logarithm of the amplitude values obtained from the DFT which is considered as the important factors for the perception of the sound.

Subsequently, the amplitude spectrum was then normalized to make further enhancements by applying Mel frequency cepstral scale. This process amplifies frequencies that are important for synthesizing human mimicking audio perception, and thus, helps the model better learn emotions in a similar manner as humans. For feature extraction, the audio signals were converted into floating-point time series and processed to generate MFCCs. These MFCCs were then transposed and averaged, providing a compact vectorial representation of each audio sample.

The final stage of the described preprocessing was the use of the trained CNN with dense layers using the extracted MFCC features. This model architecture was selected for its ability to classifiy emotions from the raw audio data, and the spatial nature of the hierarchy, as well as the temporal dimension contained in the MFCCs. In addition to promoting the accurate identification of the desired features, this methodological framework also provides the broad and accurate model of the differentiation of emotions based on the inputs of various types of sound waves.



Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	[(None, 40, 1)]	0	[]
conv1d (Conv1D)	(None, 40, 64)	256	['input_1[0][0]']
batch_normalization (BatchNormalization)	(None, 40, 64)	256	['conv1d[0][0]']
activation (Activation)	(None, 40, 64)	0	['batch_normalization[0][0]']
max_pooling1d (MaxPooling1D)	(None, 20, 64)	0	['activation[0][0]']
dropout (Dropout)	(None, 20, 64)	0	['max_pooling1d[0][0]']
conv1d_1 (Conv1D)	(None, 20, 128)	24704	['dropout[0][0]']
batch_normalization_1 (BatchNormalization)	(None, 20, 128)	512	['conv1d_1[0][0]']
activation_1 (Activation)	(None, 20, 128)	0	['batch_normalization_1[0][0]']
conv1d_3 (Conv1D)	(None, 20, 128)	24704	['dropout[0][0]']
conv1d_2 (Conv1D)	(None, 20, 128)	49280	['activation_1[0][0]']
batch_normalization_3 (BatchNormalization)	(None, 20, 128)	512	['conv1d_3[0][0]']
batch_normalization_2 (BatchNormalization)	(None, 20, 128)	512	['conv1d_2[0][0]']
add (Add)	(None, 20, 128)	0	['batch_normalization_3[0][0]', 'batch_normalization_2[0][0]']
activation_2 (Activation)	(None, 20, 128)	0	['add[0][0]']
max_pooling1d_1 (MaxPooling1D)	(None, 10, 128)	0	['activation_2[0][0]']
dropout_1 (Dropout)	(None, 10, 128)	0	['max_pooling1d_1[0][0]']
conv1d_4 (Conv1D)	(None, 10, 256)	98560	['dropout_1[0][0]']
batch_normalization_4 (BatchNormalization)	(None, 10, 256)	1024	['conv1d_4[0][0]']
activation_3 (Activation)	(None, 10, 256)	0	['batch_normalization_4[0][0]']
batch_normalization_6 (BatchNormalization)	(None, 10, 256)	1024	['conv1d_6[0][0]']
batch_normalization_5 (BatchNormalization)	(None, 10, 256)	1024	['conv1d_5[0][0]']
add_1 (Add)	(None, 10, 256)	0	['batch_normalization_6[0][0]', 'batch_normalization_5[0][0]']
activation_4 (Activation)	(None, 10, 256)	0	['add_1[0][0]']
max_pooling1d_2 (MaxPooling1D)	(None, 5, 256)	0	['activation_4[0][0]']
dropout_2 (Dropout)	(None, 5, 256)	0	['max_pooling1d_2[0][0]']
conv1d_7 (Conv1D)	(None, 5, 512)	393728	['dropout_2[0][0]']
batch_normalization_7 (BatchNormalization)	(None, 5, 512)	2048	['conv1d_7[0][0]']
activation_5 (Activation)	(None, 5, 512)	0	['batch_normalization_7[0][0]']
conv1d_9 (Conv1D)	(None, 5, 512)	393728	['dropout_2[0][0]']
conv1d_8 (Conv1D)	(None, 5, 512)	786944	['activation_5[0][0]']
batch_normalization_9 (BatchNormalization)	(None, 5, 512)	2048	['conv1d_9[0][0]']
batch_normalization_8 (BatchNormalization)	(None, 5, 512)	2048	['conv1d_8[0][0]']
add_2 (Add)	(None, 5, 512)	0	['batch_normalization_9[0][0]', 'batch_normalization_8[0][0]']
activation_6 (Activation)	(None, 5, 512)	0	['add_2[0][0]']
max_pooling1d_3 (MaxPooling1D)	(None, 2, 512)	0	['activation_6[0][0]']
dropout_3 (Dropout)	(None, 2, 512)	0	['max_pooling1d_3[0][0]']
global_average_pooling1d (GlobalAveragePooling1D)	(None, 512)	0	['dropout_3[0][0]']
dense (Dense)	(None, 512)	262656	['global_average_pooling1d[0][0]']
dropout_4 (Dropout)	(None, 512)	0	['dense[0][0]']
dense_1 (Dense)	(None, 256)	131328	['dropout_4[0][0]']
dropout_5 (Dropout)	(None, 256)	0	['dense_1[0][0]']
dense_2 (Dense)	(None, 8)	2056	['dropout_5[0][0]']

Total params: 2474376 (9.44 MB)
Trainable params: 2468872 (9.42 MB)
Non-trainable params: 5504 (21.50 KB)

Fig. 1. Detailed Overview of the Proposed Classifier's Architecture

The architecture of the deep neural network employed for the classification task is depicted operationally in Figure 1.The methodology implemented for training the emotion classification model involved constructing a deep neural network to process 40-feature input vectors derived from audio files, each encapsulating a 2-second frame of speech. These features passed through a 1D CNN with the layers activated by ReLU and dropout in the range of 20-50

IV. Model Evaluation

Model evaluation was done using the Split data set method in which the data was split into training set and test set. Subsequently, having the trained neural network, it was tested using the testing set for effectiveness and ability to perform well on data that it had not been trained on. Policies were learned for a fixed number of epochs, and performance metrics including loss and accuracy were plotted to monitor

improvements over training iterations. Furthermore, the model was again tested for accuracy using confusion matrix and classification report which gives precise measures of precision, recall, the F1 measure and the accuracy of each emotion class. They pointed out that this kind of evaluation greatly assisted in making the model strong and fast in its task of classifying emotions from audio data.

Dataset: The dataset used in this research is composed by Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and Toronto Emotional Speech Set(TESS). The RAVDESS dataset comprises of both the audio and video recording of the 24 professional actors (12 Female, and 12 Male) reciting 51 4-syllable long lexically matched Neutral North American English statements. For this study, only the audio which was converted into speech files was chosen and used for analysis Total recordings we have in our data base were 1440. These recordings cover eight emotions, namely, calm, happy, sad, angry, fearful, surprise, and disgust in two levels of emotional engagement and a neutral expression. In RAVDESS, the file naming consists of a numerical part that includes 7 numbers that systematically encode the modality, vocal channel, emotion, emotional intensity, statement, repetition, and actor number.

e.g., 03-01-06-01-02-01-12.wav

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). NOTE: There is no strong intensity for the 'neutral' emotion.

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).

On the other hand, TESS consists of recordings of two actresses reading seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral) in the carrier phrase *"Say the word. "* This generous dataset consists of 2800 files: 400 files for each of the emotions.

For this study, the RAVDESS and TESS databases were combined to increase both the variety and the amount of data in the training set, which improved the model's ability to control for variation in emotions and their intensity. The indexed list has 592 files for each of the four emotion classes, while for the neutral class, there are only 496 files because the number of neutral recordings in RAVDESS is not balanced. For integration purposes, new folders with labels as Actor25 for young adult female (YAF) from TESS and Actor26 for older adult female (OAF) voices were created. The audio files obtained from TESS were then duplicated and renamed according to the structure used in the RAVDESS dataset, numbering each file based on the emotion category it characterized. Furthermore, additional files of calm emotions were created by copying files and changing their tags to depict a calm mood, which also increased the richness of the dataset. The above structure made it easier to classify the data and come up with a broad range of compiled emotional speech data to be used in the classification of emotions.

V. Discussion of Results

The evaluation results [Table I] from our emotion classification model demonstrate high performance across all measured metrics, particularly showcasing the model's capability to maintain a consistent balance between precision and recall, as illustrated in the performance summary. More specifically, the model worked very well for the 'Calm' emotion, with a high precision of 0.94, a perfect recall of 1.00, and thus an F1 score of 0.97. This indicates very good ability to identify and classify calm emotions correctly. Although most emotions are well-classified with F1-scores around 0.94, the emotions 'Sad' and 'Fearful' have slightly lower F1-scores of 0.91 and 0.92, respectively, in agreement with known difficulties due to the literature of emotion recognition because of subtle cues for expression. Overall high scores for consistency across all emotions underline the robustness and efficiency of our model, which would be helpful in defining a really large range of emotional states from audio data.

Emotion	Precision	Recall	F1-score	Support
0-Neutral	0.95	0.94	0.95	342
1-Calm	0.94	1.00	0.97	360
2-Happy	0.94	0.90	0.92	351
3-Sad	0.92	0.90	0.91	413
4-Angry	0.94	0.96	0.95	413
5-Fearful	0.91	0.94	0.92	387
6-Disgust	0.96	0.93	0.94	320
7-Surprised	0.95	0.94	0.95	320
Accuracy	-	-	0.94	2906
Macro avg	0.94	0.94	0.94	2906
Weighted avg	0.94	0.94	0.94	2906

TABLE I
PERFORMANCE METRICS FOR EMOTION CLASSIFICATION MODEL

Furthermore, the performance of our CNN-MFCC model, as illustrated in the table [Table II], demonstrates substantial improvements in emotion classification accuracy compared to previous studies employing different methods such as MLP and SVM. Our model outperforms the MLP across all comparable emotions, with notable enhancements in the 'Happy' and 'Neutral' categories where the F1-scores have increased significantly from 0.517 and 0.411 to 0.92 and 0.95,

respectively. Compared to the SVM approaches reported by Iqbal et al., 2019 and Zhang et al., 2016, our CNN-MFCC model also shows superior or competitive results. For instance, while the SVM [15]achieved a perfect score in 'Angry', our model closely follows with an F1-score of 0.95. Furthermore, in categories such as 'Calm' and 'Fearful', our model surpasses the SVM benchmarks significantly, demonstrating the effectiveness of employing CNNs combined with MFCC for nuanced auditory emotion recognition. These comparisons underline the efficacy of our model in capturing and classifying emotional nuances more accurately than previously established models.

Class	MLP	[19] SVM	[20] SVM	CNN-MFCC
Angry	0.698	1.0	0.86	0.95
Happy	0.517	0.66	0.92	0.92
Neutral	0.411	0.93	0.76	0.95
Sad	0.685	0.67	0.71	0.91
Calm	0.759	-	0.88	0.97
Fearful	0.727	-	0.86	0.92
Disgusted	-	-	-	0.95
Surprised	-	-	0.71	0.88

TABLE II
COMPARISON OF EMOTION CLASSIFICATION METHODS

Building on the high F1-scores and comparative advantages highlighted in the previous discussions [Table I and Table II], the fine-tuning process of our model further emphasizes its robustness and stability throughout the training phases, as detailed in the accompanying graphs (Fig 2 and Fig 3). The accuracy graph shows a stable increase in both training and test accuracies, converging to an optimal level without significant discrepancies, suggesting effective generalization capabilities. Similarly, the loss graph illustrates a consistent decrease in both training and test errors, stabilizing after about 100 epochs. This convergence of loss and accuracy metrics not only complements the earlier discussions on precision, recall, and F1 scores but also underscores the model's ability to maintain high performance without overfitting, across a diverse set of emotional classes. This seamless integration of model tuning and performance metrics bolsters the confidence in the model's utility for real-world applications, where reliable and accurate emotion classification is crucial.

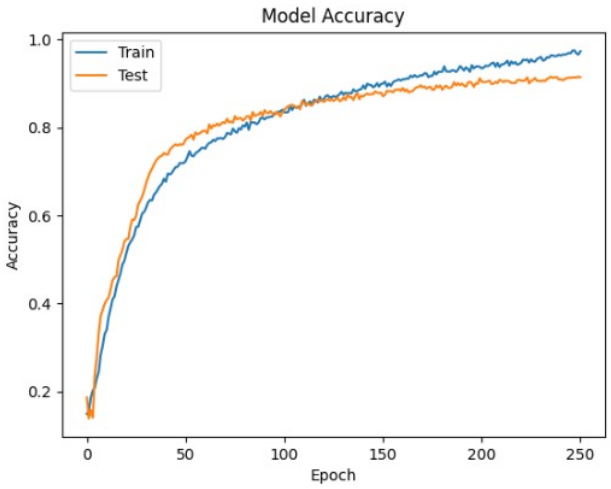


Fig. 2. Model Accuracy Over Epochs

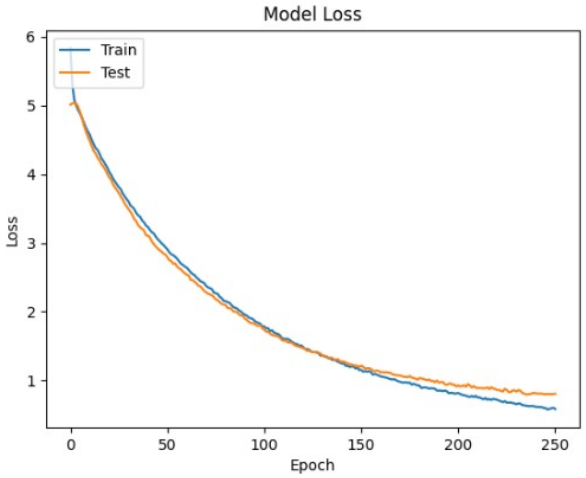


Fig. 3. Model Loss Over Epochs

VI. CONCLUSION

We have shown in this paper how the CNN model with MFCC feature extraction can perform good or fair in recognizing emotions from the audio data. Therefore, this work benefits the field of SER noticeably since it focuses on a major drawback that has been faced in the past by other models which is the variability of the dataset and its generalization problems. Combining the RAVDESS and TESS databases for the training of the model means that he learned to process and react to a more diverse selection of stimuli, hence effectively applying to different extents of the emotion and varied voice modulation. This enhanced the model of accuracy plus generalization to other unseen data sets making the entire model a potential candidate to be used in conditions where reliable emotion detection is of paramount importance. Our model has achieved an F1-score of approximately 0. Hence, the model achieves 0.94 on the test set, outstanding 'Calm' emotion, and the acceptance of other sentiments at par. This suggests a high state of versatility for detecting small emotional signals in a voice, which is imperative among use cases that involve voice assistants to diagnose. The high performance and the fact that the accuracy is sustained across all datasets independently of the type of data indicates the utility of CNNs in the promotion of SER technologies. This work provides future research with theoretical frameworks to build on for more enhancements in the technologies used in emotion detection and incorporation of the concepts into different human-computer interfaces.

REFERENCES

[1] HAYNES, J.-D., AND REES, G. Neuroimaging: decoding mental states from brain activity in humans. Nature Reviews Neuroscience 7, 7 (2006), 523.

[2] EKMAN, P. Basic emotions. Handbook of cognition and emotion 98, 45-60 (1999), 16.

[3] LOGAN, B., ET AL. Mel frequency cepstral coefficients for music modeling. In ISMIR (2000), vol. 270, pp. 1–11.

[4] F. Güneş Eriş and E. Akbal, "Enhancing speech emotion recognition through deep learning and handcrafted feature fusion," Appl. Acoust., vol. 222, p. 110070, Jun. 2024, doi: 10.1016/j.apacoust.2024.110070.

[5] S. Aruna, G. Usha, A. Saranya, M. Maheswari, and M. Annapoorna Sai Sriram Mandalika, "Deep Learning-Based Speech Emotional Analysis Using Convolution Neural Network: Bi-Directional Long Short-Term Memory," in Advances in Psychology, Mental Health, and Behavioral Studies, M. Rai and J. K. Pandey, Eds., IGI Global, 2024, pp. 96–116. doi: 10.4018/979-8-3693-4143-8.ch005.

[6] N. Bhaal, G. Usha, A. Verma, and S. Aruna, "Deep Learning for Emotion Detection in Speech: Techniques and Applications," in Advances in Psychology, Mental Health, and Behavioral Studies, M. Rai and J. K. Pandey, Eds., IGI Global, 2024, pp. 117–148. doi: 10.4018/979-8-3693-4143-8.ch006.

[7] A. Wurst, M. Hopwood, S. Wu, F. Li, and Y.-D. Yao, "Deep Learning for the Detection of Emotion in Human Speech: The Impact of Audio Sample Duration and English versus Italian Languages," in 2023 32nd Wireless and Optical Communications Conference (WOCC), Newark, NJ, USA: IEEE, May 2023, pp. 1–6. doi: 10.1109/WOCC58016.2023.10139686.

[8] L. W. Chew, K. P. Seng, L.-M. Ang, V. Ramakonar, and A. Gnanasegaran, "Audio-Emotion Recognition System Using Parallel Classifiers and Audio Feature Analyzer," in 2011 Third International Conference on Computational Intelligence, Modelling & Simulation, Langkawi, Malaysia: IEEE, Sep. 2011, pp. 210–215. doi: 10.1109/CIMSim.2011.44.

[9] N. Morales, L. Gu, and Y. Gao, "Adding noise to improve noise robustness in speech recognition," in Interspeech 2007, ISCA, Aug. 2007, pp. 930–933. doi: 10.21437/Interspeech.2007-335.

[10] C. Demiroglu and D. V. Anderson, "Noise robust digit recognition with missing frames," in 8th European Conference on Speech Communication and Technology (Eurospeech 2003), ISCA, Sep. 2003, pp. 2165–2168. doi: 10.21437/Eurospeech.2003-509.

[11] N. N. Doshi, M. U. Maniyar, K. K. Shah, N. D. Sarda, M. Narvekar, and D. Mukhopadhyay, "A Convolutional Recurrent Neural Network-Based Model For Handwritten Text Recognition To Predict Dysgraphia," in 2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS), Coimbatore, India: IEEE, Feb. 2023, pp. 145–150. doi: 10.1109/ICISCoIS56541.2023.10100514.

[12] F. A. Machot, A. H. Mosa, A. Fasih, C. Schwarzlmüller, M. Ali, and K. Kyamakya, "A Novel Real-Time Emotion Detection System for Advanced Driver Assistance Systems," in Autonomous Systems: Developments and Trends, vol. 391, H. Unger, K. Kyamaky, and J. Kacprzyk, Eds., in Studies in Computational Intelligence, vol. 391. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 267–276. doi: 10.1007/978-3-642-24806-1_21.

[13] C. Barhoumi and Y. B. Ayed, "Real-Time Speech Emotion Recognition Using Deep Learning and Data Augmentation," May 03, 2023. doi: 10.21203/rs.3.rs-2874039/v1.

[14] C. Bhanuprakash and H. Sushma, "Identification of Emotions in a Given Speech Audio Clip by Using ANN," in 2024 International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications (ICETCS), Bengaluru, India: IEEE, Apr. 2024, pp. 1–6. doi: 10.1109/ICETCS61022.2024.10543464.

[15] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech Emotion Recognition Using CNN," in Proceedings of the 22nd ACM international conference on Multimedia, Orlando Florida USA: ACM, Nov. 2014, pp. 801–804. doi: 10.1145/2647868.2654984.

[16] X. Lin, J. Liu, and X. Kang, "Audio Recapture Detection With Convolutional Neural Networks," IEEE Trans. Multimed., vol. 18, no. 8, pp. 1480–1487, Aug. 2016, doi: 10.1109/TMM.2016.2571999.

[17] F. Chenchah and Z. Lachiri, "Impact of gender and emotion type in dialogue emotion recognition," in 2014 1st International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Sousse, Tunisia: IEEE, Mar. 2014, pp. 464–467. doi: 10.1109/ATSIP.2014.6834656.

[18] MUDA, L., BEGAM, M., AND ELAMVAZUTHI, I. Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. arXiv preprint arXiv:1003.4083 (2010).

[19] IQBAL, A., AND BARUA, K. A real-time emotion recognition from speech using gradient boosting. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) (2019), IEEE, pp. 1–5.

[20] ZHANG, B., ESSL, G., AND PROVOST, E. M. Recognizing emotion from singing and speaking using shared models. In 2015 International Conference on Affective Computing and Intelligent Interaction (ACII) (2015), IEEE, pp. 139–145.