2025 2nd International Conference on Trends in Engineering Systems and Technologies (ICTEST) | 979-8-3315-0537-0/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICTEST64710.2025.11042455

Audio-Based Emotion Analysis Using Deep Learning Techniques

Siji Jose Pulluparambil
Department of Artificial Intelligence and Data Science
Adi Shankara Institute of Engineering and Technology
Kalady, India
sijijohn2223@gmail.com

Akshay S
Department of Artificial Intelligence and Data Science
Adi Shankara Institute of Engineering and Technology
Kalady, India
akshays12d@gmail.com

Don Sanson
Department of Artificial Intelligence and Data Science
Adi Shankara Institute of Engineering and Technology
Kalady, India
donsanson03@gmail.com

Fernando J Thiruthanathil
Department of Artificial Intelligence and Data Science
Adi Shankara Institute of Engineering and Technology
Kalady, India
fernandojt2003@gmail.com

Gokul Kiran R
Department of Artificial Intelligence and Data Science
Adi Shankara Institute of Engineering and Technology
Kalady, India
gokul.kiran03@gmail.com

***Abstract*—This paper uses audio parameters such as MFCCs, zero-crossing rate, chroma features, RMS values, and Mel spectrograms to provide a novel approach to machine learning for speech emotion recognition. Heterogeneous audio recordings from datasets RAVDESS, CREMA-D, TESS, and SAVEE are used for the experiment. Data augmentation of the dataset was done by performing pitch modification, stretching, shifting, and addition of noise in order to improve model robustness. CNN, being competitive with regard to emotion categorization and sequential data analysis, is being explored for possible application to sentiment analysis, mental health monitoring, or even human-computer interaction; the work, thus, makes it possible for advancing a scalable, intuitive, yet accurate system in speech emotion recognition.**

***Index Terms*—Speech emotion recognition, MFCCs, CNN, data augmentation, emotion classification, audio processing, human-computer interaction.**

I. Introduction

Voice Emotion Recognition (VER) is an important research topic in the broader subject of affective computing, which aims to allow computers to perceive, understand, and respond to human emotions. There are numerous applications for accurately detecting emotions in speech recordings, such as sentiment analysis, virtual assistants, human-computer interaction, and mental health monitoring. In-depth research and implementation of machine learning techniques for a voice emotion recognition system are covered in this study.

Emotions play a significant role in human communication and interaction and determine our choices, behavior, and relationships with others. Traditionally, the tone of voice, facial expressions, and body language have been the primary methods for communicating emotions. However, the increased reliance on digital communication platforms and the development of speech enabled technology make automatic emotion recognition from voice data imperative.

The objective of our research is to design an accurate and reliable vocal emotion detection system that can identify and classify a range of emotional states, including surprise, calm, anger, fear, sadness, and happiness. To achieve this, we use machine learning techniques, specifically deep learning models, which automatically extract relevant features from raw data.

Our experiment covers a large variety of datasets, such as the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), the Crowd Emotion Database (CREMA-D), the Toronto Emotional Speech Set (TESS), and the Surrey Audio-Visual Expressed Emotion (SAVEE) dataset. These datasets, covering a wide range of actor presented emotional expressions, offer our model an extensive training and evaluation environment.

The process entails a few crucial steps:

1) Data Preprocessing: A wide variety of audio signal processing techniques, consisting of waveplot visualization, spectrogram analysis, and feature extraction (such as Zero-crossing rate, Mel-frequency cepstral coefficients, etc.), is used in transforming the raw audio data into a form amenable to machine learning algorithms.

2) Model Building: We develop and train a CNN architecture specifically designed to handle sequential data in our machine

learning pipeline. We complement the training data to make the model robust and performative by techniques like shifting, time stretching, noise addition, and pitch modulation.

3) Evaluation and Validation: In order to evaluate the performance of the trained model, we use standard metrics that include precision, recall, accuracy, and F1-score. We use cross-validation and have actually tested it on a different validation set to guarantee that it can be generalized. Our input into the field comprises: 1) An extensive investigation into voice emotion identification methods using several datasets.

2) Creation of a deep learning model specifically designed to classify emotions in voice recordings.

3) Creation of an intuitive web application that uses submitted audio samples to forecast emotions in real time.

4) Assessment of model performance and comparison with current VER methodologies.

In general, our work attempts to push the boundaries of vocal emotion identification technology, opening doors for applications related to mental health, emotional analytics, and improved human-computer interaction.

II. Related Works

J. Li et al. [1] gives the basic ideas and methods of speech recognition, which serve as the foundation for our VER project's processing and comprehension of audio input. Their research establishes a solid foundation in feature extraction and signal processing, which are critical for deciphering audio recordings' emotions.

J. Hook et al. [2] focus on employing paralinguistic elements for automatic speech-based emotion identification. This article is especially pertinent since it extends our strategy of extracting features such as Zero-crossing rate, Mel-frequency cepstral coefficients, and more by investigating feature extraction techniques that can extract emotional indicators from speech.

M. Chatterjee et al. [3] examine the identification of voice emotions in children with cochlear implants and their classmates who are ordinarily hearing. While our initiative does not aim to cater to any particular demographic, this study highlights the significance of taking into account a variety of user groups and potential difficulties when it comes to voice detection of emotions.

B. W. Schuller [4] provides a thorough analysis of speech emotion recognition, including current developments and benchmarks. In order to measure the efficacy of our VER system, this study offers insightful information on evaluation metrics, model performance assessment, and the development of approaches in the field.

F. W. Smith and S. Rossit [5] examine how emotions are expressed on the face, emphasizing the value of multimodal emotion identification. Although this paper emphasizes the larger context of emotion detection across several modalities, which might improve the overall comprehension of emotional states, our project concentrates on voice-based emotion recognition.

M. Chen et al. [6] demonstrate an emotion communication system while highlighting the usefulness and potential influence of emotion detection technologies. Their work aligns with our project's objective of improving human computer interaction and emotional analytics skills by creating an intuitive application for real-time emotion prediction from voice recordings.

A. Gupta and D. Mishra [7] give a study on sentimental voice recognition that focuses on exploiting vocal signals to analyze emotions. Though with a focus on sentiment analysis, this study is in line with our project's goal of identifying emotional indicators from speech recordings and enhances our more comprehensive approach to emotion recognition.

K. S. Chintalapudi et al. [8] investigate the use of deep learning algorithms for speech emotion recognition. They study sophisticated neural network architectures for speech-based emotion classification, which is similar to how we use deep learning more specifically, CNNs (Convolutional Neural Networks) in our VER system.

H. Li et al. [9] contribute to the field of deep neural network-based approach voice emotion recognition research. Their paper emphasizes the importance of deep learning models in getting reliable emotion categorization from voice data, and it covers approaches and advances comparable to those we use in our project.

I. Idris and M. S. H. Salam [10] suggests a hybrid method for emotion recognition that combines prosodic characteristics and voice quality. Although the main focus of our effort is feature extraction from unprocessed audio signals, our work also highlights hybrid methodologies and alternate feature sets that can improve the accuracy of emotion recognition.

S. N. Atkar et al. [11] provide a technique that uses a CNN classifier and conversation emotion decoder to recognize emotions in speech. This paper highlights the ongoing development of innovative architectures for emotion identification tasks, which is relevant to our project that uses CNNs for sequential data analysis and classification.

III. Methodology

A. Data Collection and Preprocessing

This paper outlines a way to develop an identification system of vocal emotion using convolutional neural networks. The proposed system utilizes data augmentation and multiple audio feature extraction methods to enhance the emotion recognition accuracy.

The dataset used in this study is a fusion of four publicly accessible datasets of emotional speech: RAVDESS, CREMA, TESS, and SAVEE. Summarized in Table I. is the data collection process and Fig 1. shows the count of emotions in the dataset.

Data preparation involves a few essential steps. These include using OS and pandas python programs to take the file path and corresponding emotions that were in audio datasets. Further, an organised pandas dataframe that was created, giving a more methodical way of managing data and storing was derived from it. Then follows a phase called data

exploration which involves using the countplot function from Seaborn to make a visualised emotional distribution on the dataset. This was an important step in identifying whether there were imbalances because uneven distribution of emotions could lead the model to being trained biased toward the majority class. This could be further improved by strategies such as oversampling or undersampling to handle this problem to get a well-balanced set of emotions for the dataset.

TABLE I
DESCRIPTION OF DATASETS USED IN EMOTION CLASSIFICATION

Dataset	Description
RAVDESS	Recordings from 24 actors expressing 7 emotions (neutral, calm, happy, sad, angry, fear, and disgust).
CREMA-D	Recordings from 7 actors expressing 7 emotions.
TESS	Recordings from 12 actors expressing 6 basic emotions (neutral, happy, sad, angry, fear, and surprise).
SAVEE	Recordings from 8 actors expressing 4 emotions (neutral, angry, happy, and sad).

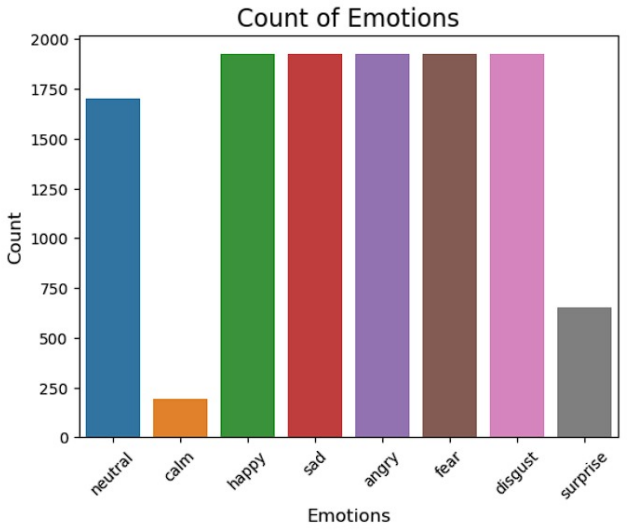

Fig. 1. Graph showing count of emotions in dataset

B. Feature Extraction

The Librosa Python module was used to perform a comprehensive feature extraction process in order to represent emotional information in audio recordings. The four key components retrieved throughout this process were the Root Mean Square (RMS) value, chroma characteristics, Zero-Crossing Rate (ZCR), and Mel-Frequency Cepstral Coefficients (MFCCs). These characteristics were picked because they have a history of accurately detecting minute emotional shifts in speech signals. Many are the methods that are utilized in feature extraction:

1) The zero-crossing rate, or ZCR, is defined as the number of times a signal crosses the zero axis in a given period. It can be computed mathematically as:

$$ZCR = \frac{\text{number of zero crossings}}{\text{signal length}} \quad (1)$$

2) MFCCs, or Mel-Frequency Cepstral Coefficients: MFCCs are the short-term power spectrum of a sound and are measured on the mel scale, a rough representation of human auditory experience. The following stages in the calculation of MFCC are a DCT, followed by logarithmic compression, and finally applying the Mel filterbank to the power spectrum of the audio data.
3) Chroma Features: Chroma features capture the distribution of energy amongst the 12 chromatic pitch classes. That is done by computing a chromagram, that is essentially a spectrogram with the frequency axis transformed in such a way to reflect the 12 chromatic pitches.
4) Root Mean Square (RMS): It is a measure of the amplitude variation of the signal over time. It can be calculated mathematically as:

$$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2} \quad (2)$$

In (2), N is the number of samples and xi represents the individual samples.

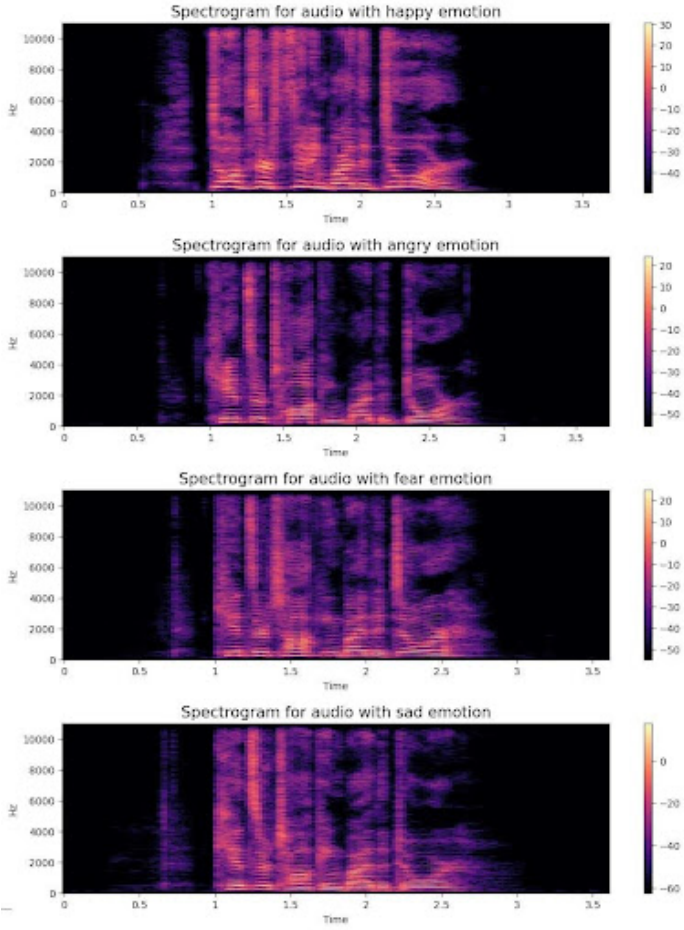

Fig. 2. Spectrogram resulting from the feature extraction process

When considered as a whole, these features extracted from the audio samples present a rich representation of the

emotional content. The spectrogram generated by the feature extraction procedure provides a visual illustration of these features, as shown in Fig. 2. Although these features have been prioritized because of their effectiveness, there is still the possibility of other variables, including pitch contours or spectral features, being considered in order to improve the functionality and degree of emotional analysis of the model.

C. Data Augmentation

In this work, input augmentation techniques were applied to increase the adaptability and resilience of the model towards variations in voice input. The techniques applied here attempted to expand and diversify the training dataset by making minor alterations to the existing audio samples. The following techniques were implemented:

1) Noise Addition. An amount of controlled white noise is added to recordings to simulate conditions of real environmental background noise for the purpose of enriching a dataset with variations in acoustic scenes.
2) Pitch Shifting: To introduce pitch changes, the audio samples' pitches were slightly raised or dropped. This modification increased the dataset's diversity by simulating different speaker characteristics and pitch variations present in actual speech data.
3) Time stretching: Playback speeds of audio samples were slightly increased or decreased upwards or downwards respectively to simulate quicker or slower speaking patterns. It resulted in an almost complete set of data incorporating the temporal displacements that always occur in normal speech.

Data augmentation is an effective technique against overfitting, which means a model has good performance on the training data but poor performance on unknown data. Fig. 3. shows the impact of various augmentation techniques on a waveform. Controlled changes ensure that the characteristics learned by the model are much more relevant in real-world settings.

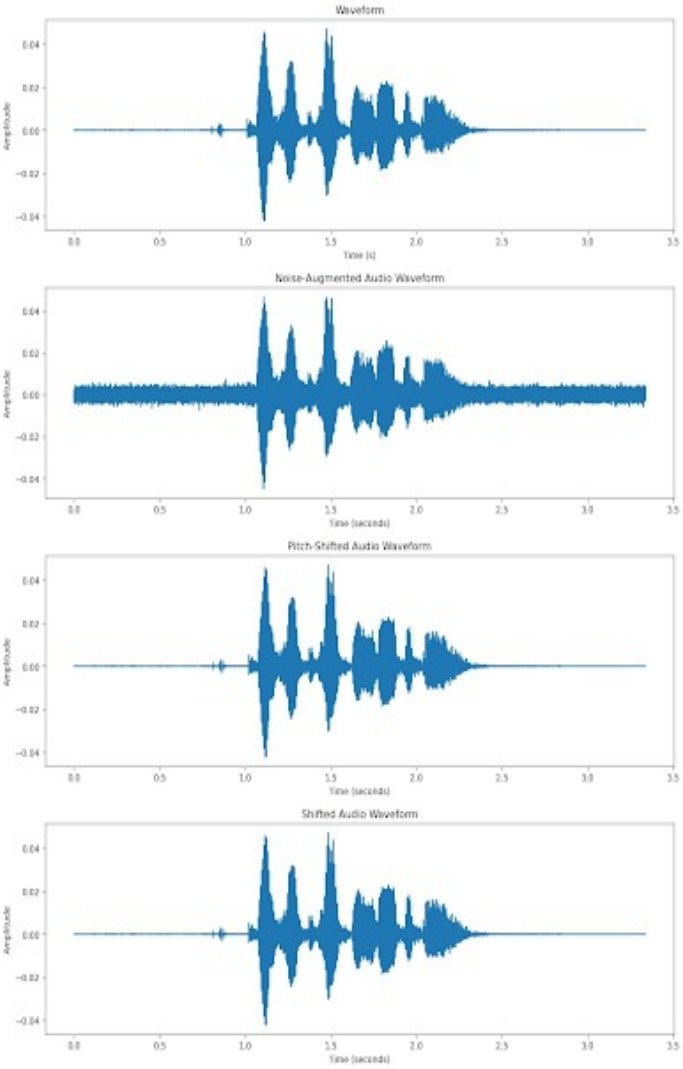

Fig. 3. Examples illustrating the impact of various augmentation techniques on a waveform

IV. Model Development

Convolutional Neural Networks (CNNs) have an inherent ability to automatically extract meaningful patterns and spatial relationships from input data. In this study, we leveraged this capability to design a CNN architecture specifically tailored for emotion recognition from voice recordings. The model was structured to effectively capture and process essential features related to vocal emotions. Fig 4. shows the general architecture of the CNN with convolutional, pooling, and fully connected layers. Below, we explain the key components that make up this architecture in detail:

1) Convolutional Layers: There are several convolutional layers, which are different from each other in the kernel sizes and activation functions adopted. The primary objective of these layers is to obtain the local features from the input audio features. In this design, the adopted ReLU (Rectified Linear Unit) activation function gives the non-linearity in the network. This first convolutional layer incorporates five as its kernel size and 256 filters.
2) Pooling Layers: The dimensionality of the data and adding invariance to feature shifts were done by including max pooling layers after every convolutional layer. For the purpose of extracting maximum values from windows of 2x2 elements along the time dimension, this pooling layer utilizes a pool size of (2, 2). It also employs the same padding for maintaining output dimensions and a stride of (2, 2) for downsampling.
3) Dropout Layers: After choosing the convolutional and dense layers, dropout layers were used in order to prevent overfitting. In this context, to enhance the learning of more resilient features a dropout rate of 20% was employed and randomly dropping neurons in training.
4) Flatten Layer : A flatten layer was used to change output from the final convolutional layer from 3D tensor to 1D vector to make the easy input into subsequent dense

layers.

5) Dense Layers: Fully connected layers at the end of the network perform classification operations by combining the information from convolutional layers for the identification of higher-level characteristics.
6) Activation Function: The softmax activation function which had been incorporated into the last dense layer had provided a generation of a probability distribution for every category of emotions.

In addition, the model was built using the Adam optimizer, a strong neural network training algorithm. As the problem of emotion recognition has a multi-class classification component, categorical cross-entropy was the loss function used in this task. Fig. 5. shows the CNN model summary with layer details and parameters.

The layered design of CNN has allowed it to correctly identify and classify emotions in audio data with the help of a number of components, such as convolutional, pooling, dropout, flatten, and dense layers.

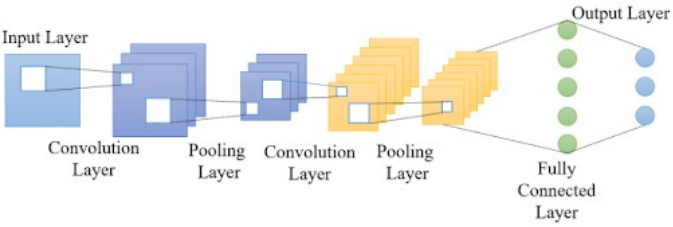

Fig. 4. CNN architecture with convolution, pooling, and fully connected layers

Model: "sequential"

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 162, 256)	1,536
max_pooling1d (MaxPooling1D)	(None, 81, 256)	0
conv1d_1 (Conv1D)	(None, 81, 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, 8)	264

Total params: 557,288 (2.13 MB)
Trainable params: 557,288 (2.13 MB)
Non-trainable params: 0 (0.00 B)

Fig. 5. CNN model summary with layer details and parameters

V. Model Training And Evaluation

The model was trained on the preprocessed and enhanced features. The training data was divided into training and validation sets using the `train_test_split` method from scikit-learn. Standard scaling on the features was applied using StandardScaler from scikit-learn to make the data normalized for improving the performance in the train.

The model was trained using a fixed number of epochs and iterations throughout the whole training dataset. Early stopping with a learning rate reduction plateau could be used to prevent overfitting. The model's performance was monitored throughout training on the validation set in order to identify the top performing model is shown below Fig. 6.

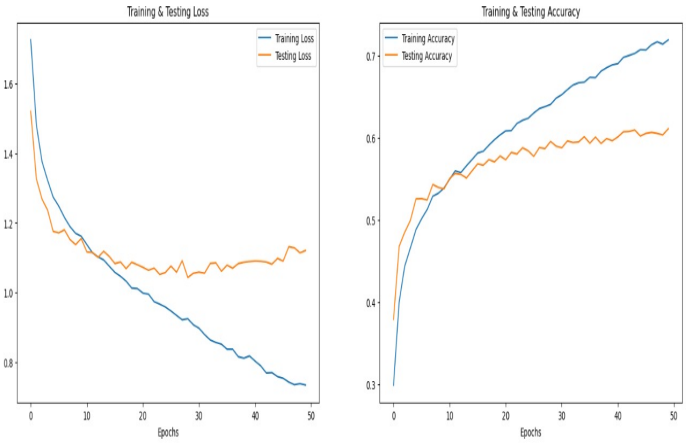

Fig. 6. Training and testing accuracy

After training, the performance of the model on an unseen test set was evaluated stringently by a variety of reliable evaluation metrics. These metrics, which are widely used and highly relevant for assessing the performance of classification models, included accuracy, precision, recall, and F1-score. A simple definition of these measures can be found below:

1) Accuracy: The percentage of correctly classified emotions is the important measure of the overall effectiveness of the model.
2) Precision: This shows how well the model classifies emotions. For a given emotion class, it is calculated as the ratio of true positives to all anticipated positives.
3) Recall: The ratio of the actual positives in the test set divided by true positives of the model measures the sensitivity in detecting instances of specific emotions.
4) F1-score: F1-score is a balanced harmonic mean of precision and recall, which balances precision and recall evaluations to provide a comprehensive view of the model's overall performance.

In addition to the quantitative evaluation criteria, the model was represented by its performance in confusion matrices towards different emotion classes. This confusion matrix represents the correct and incorrect classifications of each emotion, giving crucial insights into the model's discriminative capabilities and shortcomings. The confusion matrix obtained is shown in Fig. 8.

The structured paradigm of training and assessment improved the capacity of the model in recognizing and classify-

	Predicted Labels	Actual Labels
0	sad	disgust
1	disgust	disgust
2	angry	angry
3	disgust	disgust
4	fear	fear
5	disgust	fear
6	fear	happy
7	happy	happy
8	disgust	sad
9	neutral	sad

Fig. 7. Comparison of predicted labels with actual labels for emotion classification

ing emotional nuances in various audio datasets by reinforced robust assessment metrics and a complete confusion matrix analysis.

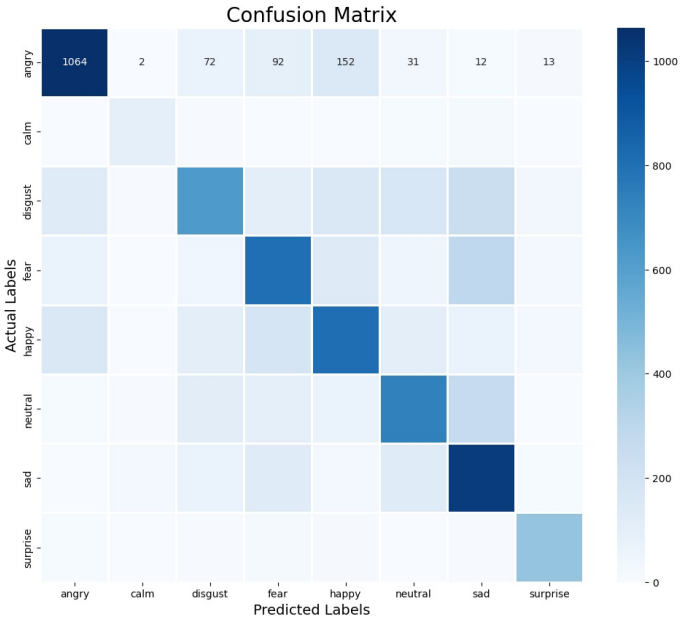

Fig. 8. Confusion Matrix

VI. Conclusion

A CNN-based voice emotion recognition system capable of extracting diverse audio features such as MFCCs, chroma, RMS, ZCR, and mel spectrogram was introduced here. Data augmentation techniques such as pitch shifting, time stretching, and noise injection were employed to make the model stronger and more generalized. These methods enabled the model to learn from varying representations of the same emotional content.

In multi-class emotion classification, the system did well, correctly classifying audio samples, especially in identifying emotions such as anger. A classification report and confusion matrix were utilized for evaluation, which revealed strengths and weaknesses in recognizing particular emotions. Delicate expressions, like sadness, were harder to identify.

By demonstrating a functional system that could be applied to areas such as sentiment analysis, customer service, and human-computer interaction, this research pushes the field of affective computing forward. In order to more accurately model the temporal nature of speech emotions, subsequent work will try out deep learning architectures such as RNNs or transformers and explore more advanced features such as GFCCs. To evaluate the impact of the model and enhance its functionality, we also expect to integrate it into real-world systems.

References

[1] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, "Fundamentals of speech recognition," in *Robust Automatic Speech Recognition: A Bridge to Practical Applications*, Waltham, MA, USA: Academic, 2016.

[2] J. Hook, F. Noroozi, O. Toygar, and G. Anbarjafari, "Automatic speech-based emotion recognition using paralinguistic features," *Bull. Polish Acad. Sci. Tech. Sci.*, vol. 67, no. 3, pp. 1–10, 2019, doi: 10.24425/bpasts.2019.129647.

[3] M. Chatterjee, D. J. Zion, M. L. Deroche, B. A. Burianek, C. J. Limb, A. P. Goren, A. M. Kulkarni, and J. A. Christensen, "Voice emotion recognition by cochlear-implanted children and their normally hearing peers," *Hearing Res.*, vol. 322, pp. 151–162, Apr. 2015, doi: 10.1016/j.heares.2014.10.003.

[4] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Commun. ACM*, vol. 61, no. 5, pp. 90–99, Apr. 2018, doi: 10.1145/3129340.

[5] F. W. Smith and S. Rossit, "Identifying and detecting facial expressions of emotion in peripheral vision," *PLoS ONE*, vol. 13, no. 5, May 2018, Art. no. e0197160, doi: 10.1371/journal.pone.0197160.

[6] M. Chen, P. Zhou, and G. Fortino, "Emotion communication system," *IEEE Access*, vol. 5, pp. 326–337, 2017, doi: 10.1109/ACCESS.2016.2641480.

[7] Gupta and D. Mishra, "Sentimental Voice Recognition: An Approach to Analyse the Emotion by Voice," *2023 International Conference on Electrical, Electronics, Communication and Computers (ELEXCOM)*, Roorkee, India, 2023, pp. 1–6, doi: 10.1109/ELEXCOM58812.2023.10370064.

[8] K. S. Chintalapudi, I. A. K. Patan, H. V. Sontineni, V. S. K. Muvvala, S. V. Gangashetty, and A. K. Dubey, "Speech Emotion Recognition Using Deep Learning," *2023 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, 2023, pp. 1–5, doi: 10.1109/ICCCI56745.2023.10128612.

[9] H. Li, X. Zhang, and M.-J. Wang, "Research on Speech Emotion Recognition Based on Deep Neural Network," *2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP)*, Nanjing, China, 2021, pp. 795–799, doi: 10.1109/ICSIP52628.2021.9689043.

[10] Idris and M. S. H. Salam, "Emotion detection with hybrid voice quality and prosodic features using Neural Network," 2014 4th World Congress on Information and Communication Technologies (WICT), Melaka, Malaysia, 2014, pp. 205–210, doi: 10.1109/WICT.2014.7076906.

[11] S. N. Atkar, R. Agrawal, C. Dhule, N. C. Morris, P. Saraf, and K. Kalbande, "Speech Emotion Recognition using Dialogue Emotion Decoder and CNN Classifier," 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2023, pp. 94–99, doi: 10.1109/ICAAIC56838.2023.10141417.