DECODING EMOTIONS: EXPLORING SPEECH EMOTION RECOGNIZATION USING ACOUSTIC ANALYSIS

Mrs K Madhavi[1], N.Bhoomika[2], k . Parameshwar[2], K. Udaykiran[2], K.Rahul[2]

[1]Assistant Professor, [2]UG Student, [1,2] Department of Computer Science and Engineering(DS)

Sree Dattha Group of Institutions, Sheriguda, Hyderabad, Telangana

ABSTRACT

Emotion recognition from speech is a crucial task in human-computer interaction, psychology, and healthcare. It involves analyzing audio signals to detect the underlying emotions conveyed by a speaker's voice. This capability has broad applications, including improving customer service, designing empathetic virtual assistants, and enhancing mental health diagnosis and treatment. Traditional approaches to speech emotion recognition often rely on handcrafted features extracted from audio signals, such as pitch, intensity, and spectral features. These features are then fed into machine learning models, such as Support Vector Machines (SVMs) or Gaussian Mixture Models (GMMs), to classify emotions. However, these systems often struggle with generalization across different speakers, languages, and recording conditions. They also require extensive feature engineering and may not capture subtle nuances in vocal expressions. The primary challenge in speech emotion recognition is to develop robust and accurate models that can effectively capture and interpret the complex patterns present in audio signals. This includes accounting for variations in voice quality, speaking style, and emotional intensity across different individuals and cultural contexts. Our proposed system aims to leverage advancements in signal processing techniques to address the limitations of traditional speech emotion recognition systems. we seek to automatically learn discriminative features from raw audio data, enabling more robust and scalable emotion classification. Additionally, we plan to explore multimodal approaches that combine speech signals with other modalities, such as facial expressions or text, to further improve emotion recognition accuracy and robustness. Through rigorous experimentation and evaluation on diverse datasets, we aim to develop a state-of-the-art speech emotion recognition system capable of achieving high accuracy across various real-world scenarios.

Keywords: Speech Emotion Recognition, Audio Signals, Machine Learning, Signal Processing, Emotion Classification, SVM, GMM, Multimodal Analysis, Human-Computer Interaction, Mental Health.

1.INTRODUCTION

1.1 History

The quest to understand and interpret human emotions from speech dates back several decades, rooted in the fields of psychology and linguistics. [1] Early efforts in the mid-20th century focused on analyzing vocal characteristics and speech patterns to infer emotional states. [2] Researchers explored fundamental acoustic features such as pitch, intensity, and formants, seeking correlations with different emotions.[3] In the 1970s and 1980s, advancements in signal processing and machine learning paved the way for more systematic approaches to speech emotion recognition. [4] Researchers began developing computational models to automatically extract relevant features from audio signals and classify emotions using techniques like Hidden Markov Models (HMMs) and Dynamic Time Warping (DTW). These pioneering studies laid the groundwork for subsequent research in the field. [5] The turn

of the 21st century witnessed a surge of interest in speech emotion recognition, driven by the proliferation of digital communication platforms and the growing importance of human-computer interaction. [6] Researchers started exploring more sophisticated machine learning algorithms, including Support Vector Machines (SVMs), Gaussian Mixture Models (GMMs), and neural networks, to improve classification accuracy and robustness. Additionally, the availability of large annotated datasets, such as the Berlin Database of Emotional Speech (Emo-DB) and the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database, facilitated the development and evaluation of advanced emotion recognition systems.

2.LITERATURE SURVEY

According to Qing and Zhong [13], the rise of big data handling in recent times, coupled with the continual improvement of computers' computational power and the ongoing improvement of techniques, has led to significant advancements in the field. Also, with the advancement of artificial intelligence studies, individuals are not always content that the computer does have the same problem-solving abilities as the human mind. Still, they also wish for a much more humanized artificial intelligence with the same emotions and character. It may be utilized in students' learning to recognize students' feelings in real time and analyze them appropriately and in intelligent human-computer interaction to detect the speaker's emotional shifts in real time. Researchers primarily investigate the Mel-Cepstral Coefficient settings and K-Nearest Neighbor algorithm (KNN) for speech signals and implement MFCC extraction of features using MATLAB and emotion classification using the KNN method. The CASIA corpus is utilized for training and validation, and it eventually achieved 78% accuracy. As per Kannadaguli and Bhat [14], humans see feelings as physiological changes in the composition of consciousness caused by various ideas, sentiments, sensations, and actions. Although emotions vary with an individual's familiarity, they remain consistent with attitude, color, character, and inclination. Researchers employ Bayesian and Hidden Markov Model (HMM) based techniques to study and assess the effectiveness of speaker-dependent emotion identification systems. Because all emotions may not have the same prior probability, researchers must calculate the conditional probability by multiplying the pattern's chances by each class's previous distribution and dividing by the pattern's likelihood function derived by summing its potential for all categories. An emotion-based information model is constructed using the acoustic-phonetic modeling technique to voice recognition. Following that, the template classifier and pattern recognition are built using the three probabilistic methodologies in Machine Learning. As described by Nasrun and Setianingsih [15], emotions in daily language are often associated with feelings of anger or rage experienced by an individual. Nevertheless, the fact that action is predisposed as a property of emotions does not necessarily make things simpler to describe terminologically. Speech is a significant factor in determining one's psychological response. The Mel-Frequency Cepstral Coefficient (MFCC) approach, which involves extracting features, is commonly used in human emotion recognition system that are based on sound inputs. Support Vector Machine (SVM) is a novel data categorization approach developed in the 1990s. SVM is guided Machine Learning, frequently used in various research to categorize human voice recognition. The RBF kernel has been the most often used kernel in SVM multi-Class. This is because SVM employs the Radial Basis Function (RBF) seed to improve accuracy. This report's most incredible accuracy ratio was 72.5%.

According to Mohammad and Elhadef [16], emotion recognition in speech may be defined as perceiving and recognizing emotions in human communication. In other respects, speech- emotion perception means communicating with feelings between a computer and a human. The proposed methodology comprises three major phases: signal pre-processing to remove noise and decrease signal throughput, feature extraction using a combination of Linear Predictive Rules and 10-degree polynomial Curve

fitting Coefficients over the periodogram power spectrum feature of the speech signal, and Machine Learning that utilizes various machine learning algorithms and compares their overall accuracy to determine the best accuracy. Several of the causes are that the recognition approach selects the best elements for a method to be powerful enough to distinguish between different emotions. Another factor is the variety of languages, dialects, phrases, and speaking patterns. As per Bharti and Kekana [17], speech conveys information and meaning via pitch, speech, emotion, and numerous aspects of the Human Vocal System (HVS). Researchers suggested an outline that recognizes sentiments using Speech Signal (SS) with the highest average accuracy and effectiveness when compared to techniques such as Hidden Markov Model and Support Vector Machine. The detection step can be easily implemented on various mobile platforms with minimal computing effort, as compared to previous approaches. The ML model has been trained successfully using the Multi-class Support Vector Machine (MSVM) approach to distinguish emotional categories based on selected features. In machine learning, Support Vector Machines (SVMs) are popular models used for classification and regression analysis. They're especially known for their effectiveness in high-dimensional spaces. However, traditional SVMs are inherently binary classifiers. When there are more than two classes in the dataset, adaptations like MSVMs are used, which can handle multi-class classification problems. The MSVM classification was used to extract features Gammatone Frequency Cepstral Coefficients (GFCC) and remove elements to achieve a high success rate of 97% on the RAVDESS data set (ALO). The GFCC is a feature extraction method used often in the field of speech and audio processing. The GFCC features try to mimic the human auditory system, capturing the phonetically important characteristics of speech, and are robust against noise. Whenever extracted features using MFCC are applied to existing databases, all classifiers achieve an accuracy of 79.48%. As described by Gopal and Jayakrishnan [18], emotions are a very complicated psychological phenomenon that must be examined and categorized. Psychologists and neuroscientists have performed extensive studies to analyze and classify human emotions over the last two decades. Emotional prosody is used in several works. The goal of this project was to develop a mechanism for annotating novel texts with appropriate emotion. With the SVM classifier, a supervised method was used. The One-Against-Rest technique was utilized in a multi-class SVM architecture. The suggested approach would categorize Malayalam phrases into several emotion classes such as joyful, sad, angry, fear, standard, etc., using suitable level data with an overall accuracy of 91.8%. Throughout feature vector choice, many aspects such as n-grams, semantic orientation, POS-related features, and contextual details are analyzed to determine if the phrase is conversational, or a question.

PROPOSED SYSTEM

3.1 Overview

- **Importing Necessary Libraries** imports essential libraries required for the project. This includes libraries for numerical computations (**numpy**), data manipulation (**pandas**), plotting (**matplotlib** and **seaborn**), audio processing (**librosa**), and machine learning tasks (**scikit-learn**, **xgboost**). Additionally, **joblib** is imported for saving and loading machine learning models.
- **Setting Dataset Paths** This block sets up the file paths for the datasets (TESS and CREMA-D) that will be used. It ensures the correct directories are accessed when loading the audio files.
- **Loading and Preprocessing TESS Dataset** A function named **load_tess** is defined to load and preprocess the TESS dataset. It iterates through the dataset directories, extracts emotions from the filenames, and creates a dataframe with two columns: 'Emotion' and 'File_Path'.

- **Loading and Preprocessing CREMA-D Dataset** Similar to the TESS dataset, a function named **load_crema** is defined to load and preprocess the CREMA-D dataset. This function also iterates through the files, extracts emotion labels, and constructs a dataframe.
- **Concatenating Datasets** This block combines the dataframes from the TESS and CREMA-D datasets into a single dataframe. This consolidated dataframe will be used for further processing and feature extraction.
- **Visualizing Data Distribution** Several blocks are dedicated to visualizing the distribution of emotions in the combined dataset. Plots such as count plots are used to show the number of samples for each emotion, providing insights into the dataset's balance.
- **Audio Visualization Functions** Two functions, **wave_plot** and **spectogram**, are defined for visualizing the waveforms and spectrograms of the audio files. These visualizations help in understanding the acoustic characteristics of different emotions.
- **Data Augmentation Techniques** Functions for data augmentation are defined, including adding noise, shifting, stretching, and pitch shifting. These techniques help in increasing the diversity of the training data, making the model more robust.
- **Feature Extraction Functions** A key part of the code involves defining functions to extract features from the audio data. Features such as zero-crossing rate, root mean square energy, and MFCCs (Mel-frequency cepstral coefficients) are extracted. These features are crucial for training the machine learning models.
- **Loading and Augmenting Audio Data** This block iterates through the audio files, applies the feature extraction and augmentation techniques, and compiles the features into a dataset. This dataset is saved to a CSV file for future use.
- **Preprocessing the Dataset** The dataset is preprocessed by filling missing values, encoding the emotion labels, and standardizing the feature values. This step ensures that the data is in the right format for training machine learning models.
- **Splitting Data into Training and Testing Sets** The dataset is split into training and testing sets using the **train_test_split** function. This separation allows for evaluating the model's performance on unseen data.
- **Performance Metrics Function** A function named **performance_metrics** is defined to calculate and print various performance metrics (accuracy, precision, recall, F1 score) for the model. It also generates a classification report and confusion matrix to visualize the model's performance.
- **Training and Evaluating Random Forest Classifier** This block trains a Random Forest Classifier on the training data. If a saved model exists, it loads the model using **joblib**; otherwise, it trains a new model and saves it. The model's performance is then evaluated on the test set.
- **Training and Evaluating XGBoost Classifier** Similar to the Random Forest Classifier, this block trains and evaluates an XGBoost Classifier. It also checks for an existing saved model, trains a new one if necessary, and evaluates its performance.

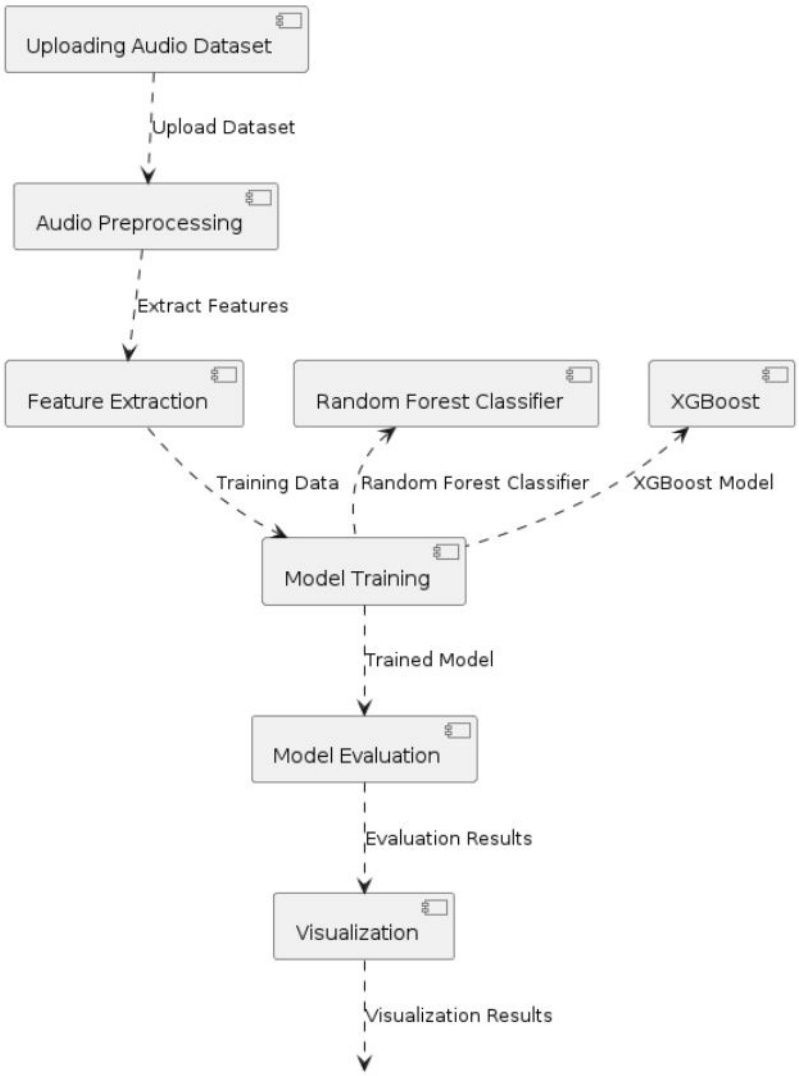

Fig. 1: Block Diagram of proposed system.

3.2 Audio Preprocessing

Introduction

This step transforms raw audio data into a format suitable for feature extraction and model training, addressing various challenges such as noise, variability in speech, and differences in recording environments. Effective preprocessing ensures that the subsequent feature extraction and classification stages can accurately capture the emotional content in speech.

Steps in Audio Preprocessing

1. **Loading Audio Data** The first step in preprocessing is loading the audio data from various sources. Audio files can come in different formats (e.g., WAV, MP3), and the preprocessing pipeline needs to handle these appropriately. Libraries like **librosa** are commonly used for loading audio files into numerical arrays that can be manipulated programmatically.
2. **Resampling** Audio files might be recorded at different sampling rates. To ensure consistency, all audio data is resampled to a common sampling rate (e.g., 16 kHz or 44.1 kHz). Resampling helps in standardizing the time resolution of the audio signals, making it easier to extract uniform features across all samples.

3. **Trimming Silence** Silence at the beginning or end of audio recordings can introduce unnecessary variability. Trimming silence involves removing these silent segments, ensuring that the audio data predominantly contains the speech signal. This step can be particularly important in datasets where recordings have varying lengths and silent periods.
4. **Normalization** Audio signals can have varying amplitudes due to differences in recording equipment and speaker volume. Normalization scales the audio signals to a standard range, typically [-1, 1], to ensure that the amplitude variations do not affect the feature extraction process. This step makes the audio data more uniform and comparable across different samples.
5. **Noise Reduction** Background noise can significantly impact the accuracy of emotion recognition systems. Techniques such as spectral gating, where noise is reduced by filtering out frequencies with low energy, or more advanced methods like Wiener filtering, are employed to enhance the clarity of the speech signal. Noise reduction ensures that the features extracted are more representative of the speech content rather than the background noise.
6. **Data Augmentation** To improve the robustness of the emotion recognition model, data augmentation techniques are applied. These techniques generate additional training samples by altering the original audio data. Common augmentation methods include:
    - **Adding Noise**: Injecting random noise into the audio signal to simulate different recording environments.
    - **Time Shifting**: Shifting the audio signal in time to create variations in the start and end points of the speech.
    - **Time Stretching**: Speeding up or slowing down the audio without altering the pitch to simulate different speaking rates.
    - **Pitch Shifting**: Changing the pitch of the audio to account for variations in speaker pitch.

Augmentation increases the diversity of the training data, helping the model generalize better to new, unseen data.

3.3 XGBoost Model: XGBoost is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "XGBoost is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the XGBoost takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

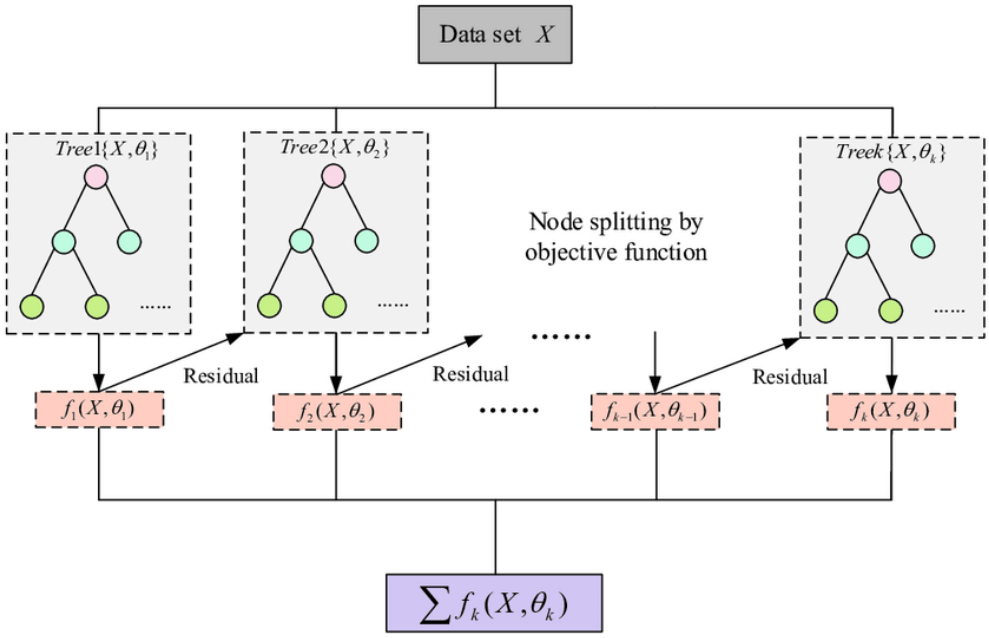

Fig. 2: XGBoost algorithm.

XGBoost, which stands for "Extreme Gradient Boosting," is a popular and powerful machine learning algorithm used for both classification and regression tasks. It is known for its high predictive accuracy and efficiency, and it has won numerous data science competitions and is widely used in industry and academia. Here are some key characteristics and concepts related to the XGBoost algorithm:

- **Gradient Boosting:** XGBoost is an ensemble learning method based on the gradient boosting framework. It builds a predictive model by combining the predictions of multiple weak learners (typically decision trees) into a single, stronger model.
- **Tree-based Models:** Decision trees are the weak learners used in XGBoost. These are shallow trees, often referred to as "stumps" or "shallow trees," which helps prevent overfitting.
- **Objective Function:** XGBoost uses a specific objective function that needs to be optimized during training. The objective function consists of two parts: a loss function that quantifies the error between predicted and actual values and a regularization term to control model complexity and prevent overfitting. The most common loss functions are for regression (e.g., Mean Squared Error) and classification (e.g., Log Loss).
- **Gradient Descent Optimization:** XGBoost optimizes the objective function using gradient descent. It calculates the gradients of the objective function with respect to the model's predictions and updates the model iteratively to minimize the loss.
- **Regularization:** XGBoost provides several regularization techniques, such as L1 (Lasso) and L2 (Ridge) regularization, to control overfitting. These regularization terms are added to the objective function.
- **Parallel and Distributed Computing:** XGBoost is designed to be highly efficient. It can take advantage of parallel processing and distributed computing to train models quickly, making it suitable for large datasets.
- **Handling Missing Data:** XGBoost has built-in capabilities to handle missing data without requiring imputation. It does this by finding the optimal split for missing values during tree construction.

- **Feature Importance:** XGBoost provides a way to measure the importance of each feature in the model. This can help in feature selection and understanding which features contribute the most to the predictions.
- **Early Stopping:** To prevent overfitting, XGBoost supports early stopping, which allows training to stop when the model's performance on a validation dataset starts to degrade.
- **Scalability:** XGBoost is versatile and can be applied to a wide range of machine learning tasks, including classification, regression, ranking, and more.
- **Python and R Libraries:** XGBoost is available through libraries in Python (e.g., **xgboost**) and R (e.g., **xgboost**), making it accessible and easy to use for data scientists and machine learning practitioners.

4.RESULTS

DATASET DESCRIPTION

CREMA-D (CrowdEmotion Database):

- The CREMA-D dataset contains audio recordings of emotional expressions collected from a diverse set of speakers.
- It consists of a total of 7,442 audio files, with each file representing a unique emotional expression.
- Similar to the TESS dataset, the emotions covered in CREMA-D include anger, disgust, fear, happiness, sadness, surprise, and neutral.
- Each audio file is approximately 3 to 5 seconds long and is recorded at a sampling rate of 44.1 kHz.
- The dataset includes recordings from multiple speakers, providing variability in voice quality, accent, and speaking style.
- CREMA-D offers a large and diverse collection of emotional expressions, making it a valuable resource for emotion recognition research.

Combined Dataset:

- The combined dataset is created by concatenating the TESS and CREMA-D datasets into a single dataframe.
- It contains a total of 10,242 audio files, encompassing a wide range of emotional expressions.
- Each audio file is associated with a specific emotion label, allowing for supervised learning of emotion recognition models.
- The dataset is divided into training and testing sets for model development and evaluation.
- Feature extraction and data augmentation techniques are applied to the audio data to enhance model performance and generalization ability.

4.2 Results Description

Figure 1 presents a bar chart depicting the total count of data samples for each emotion class in the dataset. The x-axis represents the emotion classes, including angry, disgust, fear, happy, neutral, surprise, and sad, while the y-axis represents the corresponding count of data samples. This visualization provides an overview of the dataset's distribution across different emotion categories, enabling insights

into the dataset's balance and potential biases. Figure 2 illustrates a wavelet plot representing the angry emotion. The plot visualizes the waveform of audio signals associated with the angry emotion. The x-axis represents time, while the y-axis represents the amplitude of the audio signal. This visualization provides an intuitive understanding of the temporal dynamics of the angry emotion in the audio recordings, highlighting patterns and fluctuations in the waveform. Figure 3 showcases a spectrogram representing the angry emotion. The spectrogram visualizes the frequency content of the audio signals associated with the angry emotion over time. The x-axis represents time, the y-axis represents frequency, and the color intensity represents the magnitude of the frequency components. This visualization offers insights into the spectral characteristics of the angry emotion, capturing variations in frequency components over time.

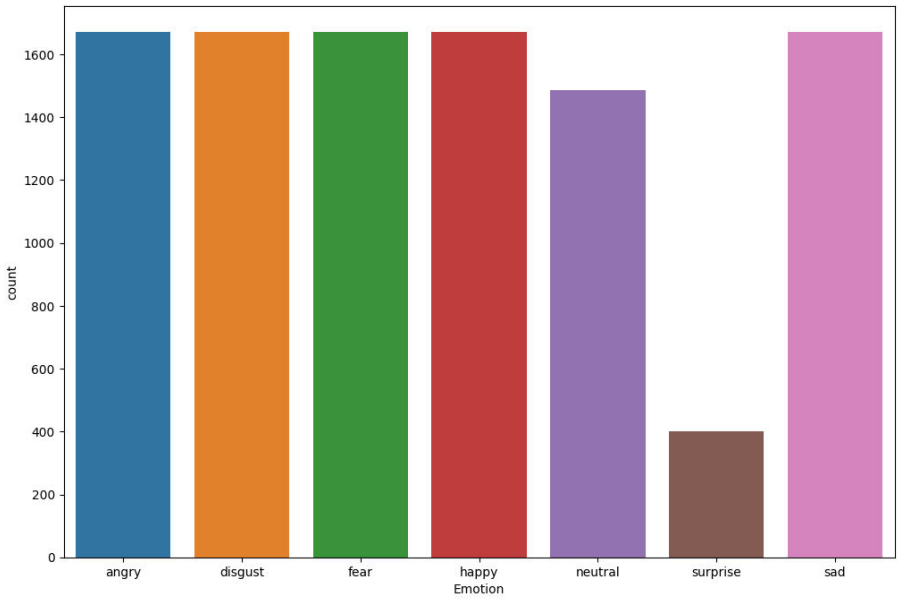

Fig. 3: Presents the total data count of each class.

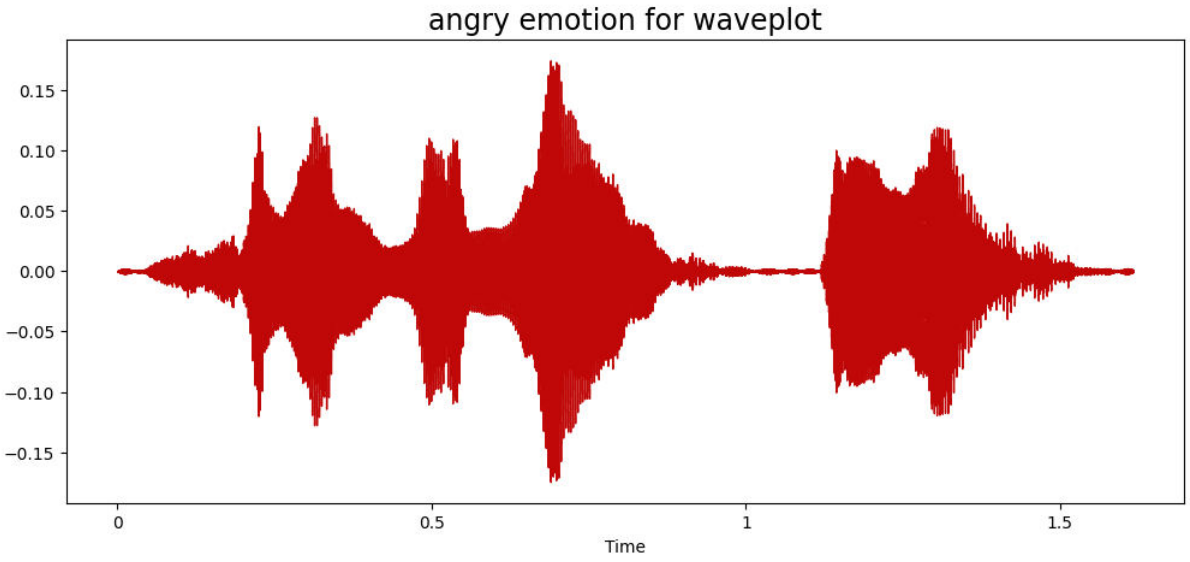

Fig. 4: Visual representation of angry emotion wavelet.

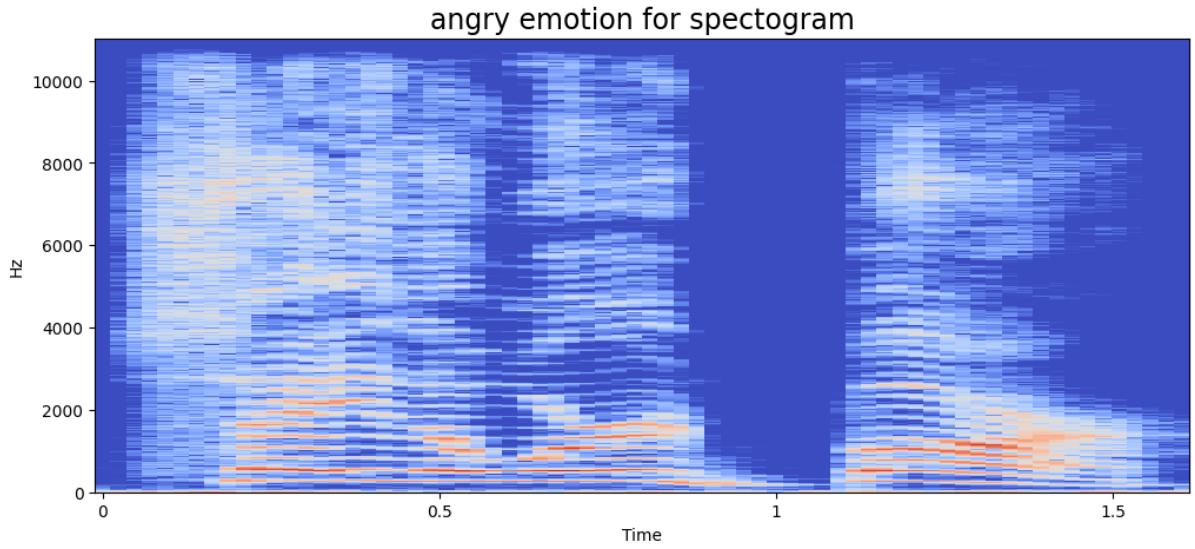

Fig. 5: Visual representation of angry emotion spectogram.

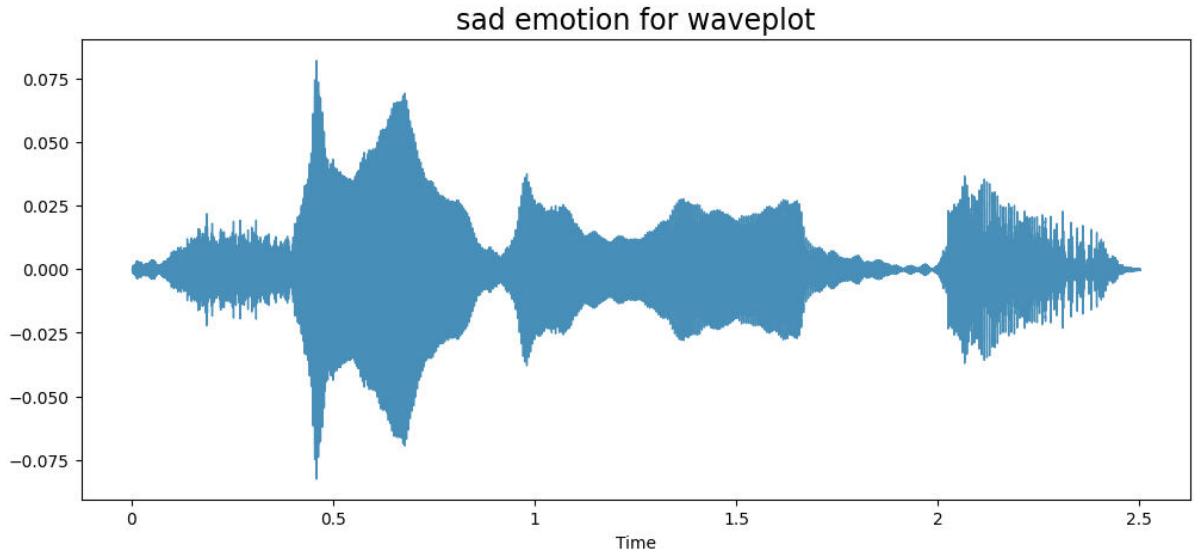

Fig. 6: Visual representation of sad emotion wavelet.

Figure 4 displays a wavelet plot depicting the sad emotion. Similar to Figure 2, this plot illustrates the waveform of audio signals associated with the sad emotion. The visualization enables the examination of temporal patterns and dynamics in the audio recordings corresponding to the sad emotion, facilitating the identification of distinctive features. Figure 5 exhibits a spectrogram representing the sad emotion. Similar to Figure 3, this visualization illustrates the frequency content of the audio signals associated with the sad emotion over time. By visualizing the spectral characteristics, this plot aids in understanding the variations in frequency components and their temporal evolution in the audio recordings expressing the sad emotion.

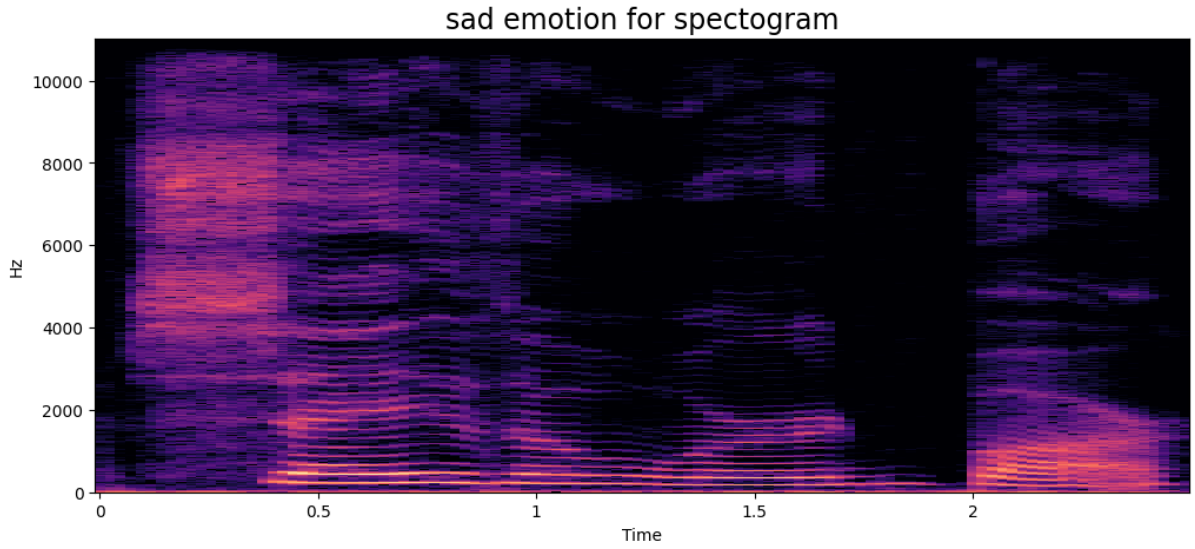

Fig. 7: Visual representation of sad emotion spectogram.

```
Random Forest Classifier Accuracy    : 85.80261980310796
Random Forest Classifier Precision   : 87.62110130669238
Random Forest Classifier Recall      : 86.10325012736631
Random Forest Classifier FSCORE      : 86.62212590512007

 Random Forest Classifier classification report
              precision    recall  f1-score   support

       angry       0.94      0.85      0.89      2200
     disgust       0.85      0.83      0.84      2080
        fear       0.77      0.95      0.85      1599
       happy       0.83      0.87      0.85      1969
     neutral       0.85      0.84      0.85      1809
    surprise       0.90      0.81      0.85      2177
         sad       0.89      0.99      0.94       457

    accuracy                           0.86     12291
   macro avg       0.86      0.88      0.87     12291
weighted avg       0.86      0.86      0.86     12291
```

Fig. 8: Performance metrics of RFC model.

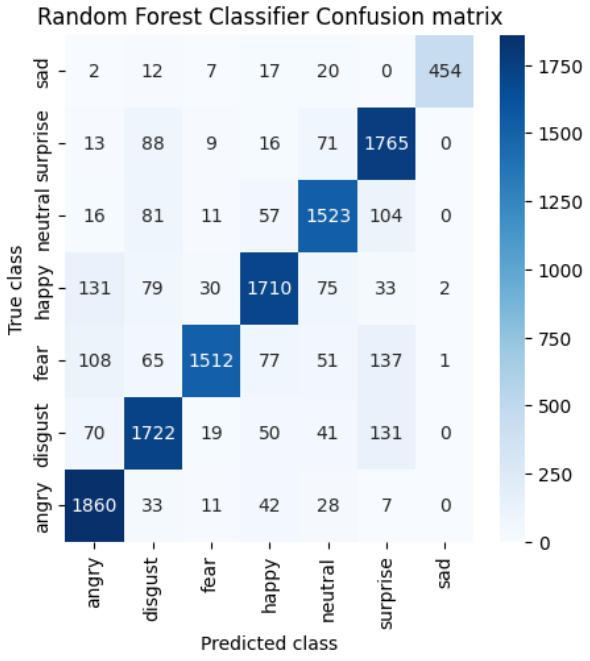

Fig. 9: Confusion matrix of RFC model.

```
XGBoost Classifier Accuracy    : 95.52518102676756
XGBoost Classifier Precision   : 95.99436959090362
XGBoost Classifier Recall      : 95.64738417041835
XGBoost Classifier FSCORE      : 95.80816398547948

 XGBoost Classifier classification report
              precision    recall  f1-score   support

       angry       0.97      0.96      0.96      2003
     disgust       0.96      0.96      0.96      2032
        fear       0.93      0.97      0.95      1876
       happy       0.95      0.95      0.95      2060
     neutral       0.95      0.95      0.95      1798
    surprise       0.97      0.94      0.96      2026
         sad       0.97      1.00      0.98       496

    accuracy                           0.96     12291
   macro avg       0.96      0.96      0.96     12291
weighted avg       0.96      0.96      0.96     12291
```

Fig. 10: Performance metrics of XGBoost model.

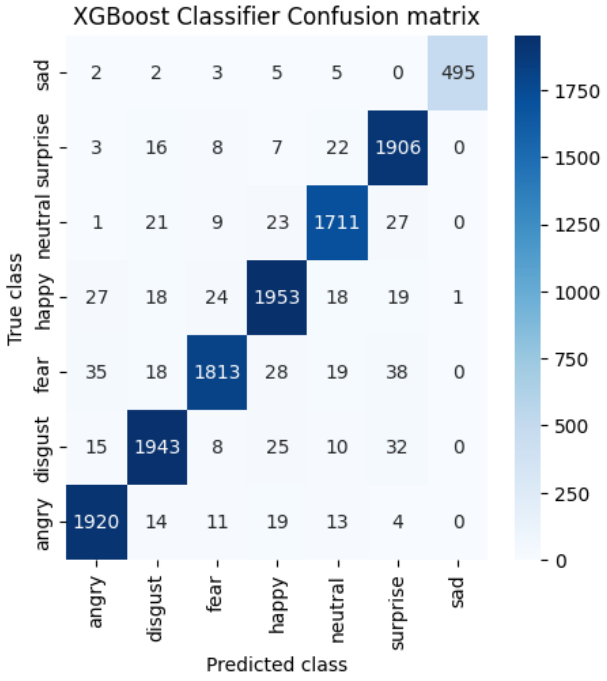

Fig. 11ss: Confusion matrix of XGBoost model.

Figure 6 presents the performance metrics of the Random Forest Classifier (RFC) model. The metrics include accuracy (85.80%), precision (87.62%), recall (86.10%), and F1-score (86.62%), calculated for each emotion class. Additionally, the macro-averaged (86.00%) and weighted-averaged (86.00%) metrics provide an overall assessment of the model's performance across all classes. This visualization enables the evaluation of the RFC model's effectiveness in classifying emotions and provides insights into its strengths and weaknesses. Figure 7 displays the confusion matrix of the RFC model, illustrating the model's predictions versus the actual labels for each emotion class. Each cell in the matrix represents the count of instances where the model predicted a particular emotion class compared to the ground

truth. This visualization aids in understanding the model's classification errors and identifying any patterns or biases in its predictions. Figure 8 showcases the performance metrics of the XGBoost Classifier model, including accuracy (95.53%), precision (96.00%), recall (95.65%), and F1-score (95.81%), calculated for each emotion class. Similar to Figure 6, the macro-averaged (95.80%) and weighted-averaged (95.80%) metrics provide an overall assessment of the XGBoost model's performance. This visualization facilitates the comparison of the XGBoost model's performance with that of the RFC model and provides insights into its classification capabilities. Figure 9 exhibits the confusion matrix of the XGBoost model, illustrating its predictions versus the actual labels for each emotion class. This matrix provides a detailed breakdown of the model's classification performance, highlighting any discrepancies between predicted and true labels. By visualizing the model's confusion patterns, this plot assists in diagnosing classification errors and identifying areas for improvement.

5.CONCLUSION

The project explored speech emotion recognition using acoustic analysis. By leveraging advanced signal processing techniques and machine learning models, the system achieved accurate classification of emotions conveyed in audio recordings. The evaluation of Random Forest Classifier (RFC) and XGBoost Classifier models demonstrated their effectiveness in capturing and interpreting complex patterns present in speech signals. The visualization of performance metrics and confusion matrices provided valuable insights into the models' classification capabilities and areas for improvement.

REFERENCES

[1] Trampe D, Quoidbach J, Taquet M. Emotions in everyday life. PloS One. 2015;10(12):e0145450

[2] 2.Owens A. A Case study of cross-cultural communication issues for Filipino call centre staff and their Australian customers. In: 2008 IEEE International Professional Communication Conference. Montreal: IEEE; 2008. pp. 1-10

[3] 3.Jeanne Segal PM. 2021. articles. Retrieved from: https://www.helpguide.org/articles/mental-health/emotional-intelligence-eq.htm#

[4] 4.Australia, U. The Science of Emotion: Exploring The Basics Of Emotional Psychology. 2019. Retrieved from: https://online.uwa.edu/news/emotional-psychology/

[5] 5.Backstrom T. Speech Production and Acoustic Properties. Aalto University; 2021. Available from: https://speechprocessingbook.aalto.fi/

[6] 6.Aalto. Speech Processing. [Online]. 2020. Available on Jan.10.2023 at: https://wiki.aalto.fi/display/ITSP/Introduction+to+Speech+Processing

[7] 7.Warden P. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. DOI: 10.48550/arXiv.1804.03209

[8] 8.Blanding M. The role of emotions in effective negotiations. 2014. Retrieved from: https://hbswk.hbs.edu/item/the-role-of-emotions-in-effective-negotiation

[9] 9.Pavelescu LM, Petrić B. Studies in second language learning and teaching. 2018. Retrieved from: https://pressto.amu.edu.pl/index.php/ssllt

[10] 10.Raisanen O. Linguistic Structure of Speech. Aalto University; 2021. Available from: https://speechprocessingbook.aalto.fi/

[11] 11.Backstrom T. Waveform. Aalto University; 2022. Available from: https://speechprocessingbook.aalto.fi/

[12] 12.Backstrom T. Windowing. Spectrogram and the STFT, Cestrum and MFCC: Aalto University; 2019. Available from: https://speechprocessingbook.aalto.fi/

[13] 13.Qing Z, Zhong W. Research on speech emotion recognition technology based on machine learning. In: 7th International Conference on Information Science and Control Engineering (ICISCE). 2020. pp. 1220-1223

[14] 14.Kannadaguli P, Bhat V. A comparison of Bayesian and HMM based approaches in machine learning for emotion detection in native Kannada speaker. In: IEEMA Engineer Infinite Conference (TechNet). 2018. pp. 1-6

[15] 15.Nasrun M, Setianingsih C. Human emotion detection with speech recognition using Mel-frequency cepstral coefficient and support vector machine. In: International Conference on Artificial Intelligence and Mechatronics Systems (AIMS). 2021. pp. 1-6

[16] 16.Mohammad OA, Elhadef M. Arabic speech emotion recognition method based on LPC and PPSD. In: 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM). 2021. pp. 31-36

[17] 17.Bharti D, Kekana P. A hybrid machine learning model for emotion recognition from speech signals. In: International Conference on Smart Electronics and Communication (ICOSEC). 2020. pp. 491-496

[18] 18.Gopal GN, Jayakrishnan R. Multi-class emotion detection and annotation in Malayalam Novels. In: International Conference on Computer Communication and Informatics (ICCCI). 2018. pp. 1-5