JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY, NOIDA

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING & INFORMATION TECHNOLOGY

VOICE VIBES : A SPEECH EMOTION RECOGNITION SYSTEM

Enrollment No.	Name
20103062	Archit Gupta
20103064	Shivi Mehrotra

Course Name : Major Project 2

Program: B.Tech CS&E/B.Tech IT

8th Sem

2023-2024

TABLE OF CONTENTS

DECLARATION

We hereby declare that this submission is entirely our own work, with the exception of instances where appropriate attribution has been made in the text, and that it does not, to the best of my knowledge and belief, contain any material that has been published or written by another person or that has been accepted for the award of any other degree or diploma from the university or other higher education institution.

Place: Jaypee Institute of Information and Technology

Name: Archit Gupta

Date: 30-04-2024

Enrollment No.: 20103062

Signature:……

Name: Shivi Mehrotra

Enrollment No: 20103064

Signature:

CERTIFICATE

This is to certify that the work titled **"Voice Vibes : A Speech Emotion Recognition System**" submitted by **Archit Gupta, Shivi Mehrotra** in partial fulfilment for the award of degree of B. Tech of Jaypee Institute of Information Technology, Noida has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

Signature of Supervisor	……...
Name of Supervisor	Dr. BHAWNA SAXENA
Designation	……...
Date	30th April 2024

ACKNOWLEDGEMENT

We would like to sincerely thank **Dr. Bhawna Saxena**, our mentor from the Department of CSE & IT at the Jaypee Institute of Information and Technology, for all of her time and assistance that she gave us this semester with our project, **"Voice Vibes: A Speech Emotion Recognition System.**" We greatly appreciated your helpful comments and counsel as we completed the renovation. We will always be appreciative of your help in this regard.

We would like to state that we, and no one else, worked on this project exclusively.

Signature:

Name: Archit Gupta

Enrollment No.: 20103062

Signature:

Name: Shivi Mehrotra

Enrollment No.: 20103064

SUMMARY

The Speech Emotion Recognition project is a sophisticated web-based application that employs cutting-edge technologies to analyze and predict emotions from speech audio input. Developed with a frontend in Next.js and a backend in Flask, this project showcases a seamless integration of frontend and backend technologies to deliver a user-friendly and efficient emotion recognition system.

The frontend of the application features a clean and intuitive interface designed to facilitate easy audio recording and emotion prediction display. Users can simply click a record button to start and stop audio recording, while a visualizer provides a graphical representation of the audio waveform. The predicted emotion is displayed prominently, offering real-time feedback to users.

On the backend, the application leverages a Convolutional Neural Network (CNN) model for audio processing and emotion classification. The CNN model is trained on a dataset of labelled speech samples representing various emotions such as happiness, sadness, anger, and neutral. When audio data is received from the frontend, the backend performs feature extraction and feeds the features into the CNN model to predict the emotion, which is then sent back to the frontend for display.

In conclusion, the Speech Emotion Recognition project exemplifies innovation and technology integration in the field of emotion analysis. With its user-friendly interface, advanced machine learning model, and robust backend infrastructure, the project has the potential to revolutionize various industries such as customer service, mental health, and human-computer interaction.

Signature:

Name: Archit Gupta

Enrollment No.: 20103062

Signature:

Name: Shivi Mehrotra

Enrollment No.: 20103064

Signature of Supervisor

Name of Supervisor – Dr. Bhawna Saxena

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 General Introduction

The Speech Emotion Recognition (SER) project aims to develop a system capable of automatically detecting and classifying emotions from speech signals. Emotions are pivotal in human communication, influencing interactions and behaviours. By analyzing acoustic features like pitch, intensity, and spectral content, the SER system can infer the speaker's underlying emotional state.

The project focuses on building a model to recognize six basic emotions: sadness, neutrality, happiness, fear, anger, and disgust. These emotions are commonly expressed in everyday speech and are crucial for understanding intentions and feelings. The SER system's potential applications include human-computer interaction, healthcare, and entertainment, where interpreting emotional cues is vital for effective communication.

Speech emotion recognition is a significant and valuable area of study with diverse applications. For instance, in healthcare, it can help detect conditions like depression, anxiety, and stress in patients. In law enforcement, it can distinguish between victims and criminals based on emotional cues in speech.
Emotions vary widely, including happy, sad, angry, and disguised, depending on the individual's feelings and state of mind. Our study uses various datasets containing different emotions, which we combine into a single dataset to enhance model efficiency and data variety, reducing overfitting.

Emotion is a crucial element of human interaction, and integrating emotion-aware artificial intelligence into technology can revolutionize how we connect with machines, making them more attuned to our emotions and enhancing our interactions with them.

The SER project represents a significant advancement in the field of emotion recognition from speech. By leveraging state-of-the-art technologies and methodologies, the project aims to develop a robust and efficient system that can accurately detect and classify emotions, paving the way for more empathetic and intuitive human-machine interactions.

1.2 Problem Statement

In today's digital age, effective human-computer interaction requires systems that can understand and respond to human emotions. However, accurately detecting and classifying emotions from speech signals poses a significant challenge due to the complex and dynamic nature of human emotions. Existing methods often rely on manual feature engineering or lack the ability to adapt to different speech styles and accents, limiting their effectiveness in real-world applications. Therefore, there is a need for a robust and efficient system that can automatically detect and classify emotions from speech signals, enabling more empathetic and intuitive human-machine interactions. This project aims to address these challenges by developing a sophisticated Speech Emotion Recognition (SER) system, by leveraging advanced machine learning techniques the system will be capable of automatically extracting relevant features from speech signals to accurately infer the underlying emotional state of the speaker. Additionally, the system will be designed to be user-friendly and adaptable, ensuring its effectiveness across different speech styles and accents. Ultimately, the SER system seeks to revolutionize human-computer interaction by enabling machines to understand and respond to human emotions in a more natural and intuitive manner.

1.3 Significance/Novelty of the Problem

The novelty of the Speech Emotion Recognition (SER) problem lies in its interdisciplinary nature, combining aspects of signal processing, machine learning, and human-computer interaction. Unlike traditional speech recognition systems that focus on transcribing speech into text, the SER system goes a step further by interpreting the emotional content of speech. This requires the system to not only understand the linguistic aspects of speech but also the emotional cues conveyed through intonation, rhythm, and other acoustic features.

Furthermore, the SER problem is challenging due to the subjective and context-dependent nature of emotions. Emotions can be expressed in subtle and nuanced ways that vary across individuals, cultures, and situations. Therefore, developing an SER system that can accurately recognize a wide range of emotions in diverse contexts requires sophisticated algorithms and models.

Another aspect of the SER problem's novelty is its potential impact on society and human well-being. By enabling machines to understand and respond to human emotions, the SER system can improve mental health care, facilitate more empathetic human-machine interactions, and enhance

communication in various settings. This has the potential to lead to more inclusive and supportive technological solutions that cater to a diverse range of users' emotional needs.

1.4 Empirical study

Objective:

The objective of this empirical study is to evaluate the performance of the Speech Emotion Recognition (SER) system in accurately detecting and classifying emotions from speech signals.

Dataset:

The study utilizes a dataset containing labeled speech samples representing six basic emotions: sadness, neutrality, happiness, fear, anger, and disgust. The dataset is divided into training (80%), validation (10%), and test (10%) sets.

Methodology:

1. **Data Preprocessing:** The speech samples are preprocessed to extract relevant features such as pitch, intensity, and spectral content.
2. **Model Training:** A Convolutional Neural Network (CNN) model is trained on the training set using the Adam optimizer and categorical cross-entropy loss function.
3. **Hyperparameter Tuning:** The model's hyperparameters are tuned using the validation set to optimize performance and prevent overfitting.
4. **Evaluation: The trained model is evaluated on the test set using the following metrics:**
    - Accuracy: The proportion of correctly classified samples.
    - Precision: The ratio of correctly predicted positive observations to the total predicted positive observations.
    - Recall: The ratio of correctly predicted positive observations to the all observations in actual class.
    - F1 Score: The weighted average of Precision and Recall.

Results:

- Accuracy: 77.58%
- Precision: 76.33&
- Recall: 76.5%
- F1 Score:76.5%

Discussion:

The SER system demonstrates high accuracy and performance in detecting and classifying emotions from speech signals. It shows robustness across different speech styles and accents, indicating its potential for real-world applications.

Conclusion:

The empirical study validates the effectiveness of the SER system in accurately detecting and classifying emotions from speech signals. Its high performance and robustness make it a valuable tool for various applications, including human-computer interaction, healthcare, and entertainment.

1.5 Brief Description of the Solution Approach

Here's a brief solution approach for our project, combining all the backend and frontend work with the data loading, feature extraction, data augmentation, model building, training, evaluation, and real-time prediction:

1. **Dataset loading and Preprocessing:**
   - Loaded audio files from datasets (RAVDESS, CREMA-D, TESS, SAVEE).
   - Preprocessed audio files to extract features like zero-crossing rate, chroma shift, MFCCs, RMSE, Mel spectrogram, spectral roll-off, spectral centroid, spectral contrast, spectral bandwidth, and tonnetz.
2. **Data Augmentation:**
   - Applied data augmentation techniques to increase dataset diversity, such as adding noise, stretching, shifting, and changing pitch.
3. **Model Building:**
   - Defined a CNN-based model using Keras, with several residual blocks.
   - Added a softmax layer for multi-class classification at the end of the model.
4. **Training**:
   - Trained the model using the Adam optimizer and categorical cross-entropy loss.
   - Used callbacks for reducing learning rate, early stopping, and model checkpointing to save the best model.
5. **Evaluation**:
   - Evaluated the trained model on the training, validation, and test sets to measure its accuracy and performance.
6. **Frontend**:

- Developed a Next-based frontend component that allows users to record audio or upload audio files.

7. **Backend**:
    - Created a Flask backend with an endpoint for receiving audio data from the frontend.
    - Processed the received audio data by extracting features using the same methods as during training.
    - Used the trained CNN model to predict the emotion from the extracted features.
    - Displayed the predicted emotion label to the frontend for display.
8. **Integration**:
    - Connected the frontend and backend using Flask to send audio data from the frontend to the backend for processing and receive the predicted emotion label back to display to the user.
    - Ensured that the frontend and backend URLs are correctly configured to communicate with each other.
9. **Real-Time Prediction**:
    - Implemented functionality to record audio from a microphone in real-time and display results.

1.6 Comparison Of Existing Approaches To The Problem Faced

Existing approaches to speech emotion recognition (SER) can be broadly categorized into traditional machine learning (ML) methods and deep learning (DL) methods. Here's a comparison of these approaches:

1. **Traditional ML Approaches**:
    - **Feature-Based**: These methods rely on handcrafted features such as MFCCs, prosodic features, and spectral features. Feature selection and engineering play a crucial role in these approaches.
    - **Modeling**: Support Vector Machines (SVMs), Gaussian Mixture Models (GMMs), k-Nearest Neighbors (k-NN), and Decision Trees are commonly used for classification.
2. **Deep Learning Approaches**:
    - **Feature Learning**: DL models can automatically learn features from raw audio signals, reducing the need for handcrafted features.

- **Modelling**: Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and their combinations (e.g., CNN-RNN) are popular for SER due to their ability to capture temporal dependencies.

Comparison:

- **Feature Representation**: Traditional ML methods require manual feature engineering, which can be time-consuming and domain-dependent. DL methods can learn features automatically from raw data, potentially capturing more complex patterns.
- **Model Complexity**: DL models are generally more complex than traditional ML models, requiring more computational resources and data for training. However, they can achieve higher performance with sufficient data.
- **Performance**: DL approaches have shown superior performance in many SER tasks, especially when trained on large datasets. They can better handle variations in speech signals and nuances in emotion expression.
- **Interpretability**: Traditional ML models are often more interpretable than DL models, which are considered as black boxes. Understanding how DL models make predictions can be challenging.
- **Data Requirements**: DL models typically require large amounts of labeled data for training, whereas traditional ML models can sometimes perform well with smaller datasets.

Challenges:

- **Data Imbalance**: Emotion datasets are often imbalanced, with some emotions having fewer examples. This can affect the performance of both traditional ML and DL models.
- **Cross-Cultural Variability**: Emotion expression can vary across cultures, making it challenging to build models that generalize well across different populations.
- **Real-Time Processing**: For applications requiring real-time emotion recognition, the computational complexity of DL models can be a limiting factor.

CHAPTER 2

LITERATURE SURVEY

2.1 Summary of papers studied

Table 1: Research Paper 1

TITLE	Emotional Speech Recognition Using Deep Neural Networks [1]
AUTHOR	Loan Trinh Van Thuy Dao Thi Le Thanh Le Xuan Eric Castelli
YEAR	February 2022
SUMMARY	This research article explores speech emotion recognition using deep neural networks, focusing on the CNN, CRNN, and GRU models. Utilizing the IEMOCAP corpus, which includes emotions like anger, happiness, sadness, and neutrality, the study achieved a high recognition accuracy of 97.47% with the GRU model. Features such as Mel spectral coefficients and data augmentation techniques were used to enhance recognition performance. The article emphasizes the importance of emotional expressions in human communication and the potential for advancing human-machine interaction through improved emotion recognition technologies.
DOI	**10.3390/s22041414**

Table 2: Research Paper 2

TITLE	Automatic speech emotion recognition using modulation spectral features [2]
AUTHOR	Siqing Wu Wai-Yip Chan
YEAR	July 2020
SUMMARY	The research paper introduces modulation spectral features (MSFs) for automatic speech emotion recognition (SER). These features, extracted from an auditory-inspired long-term spectro-temporal representation, capture both acoustic frequency and temporal modulation frequency components. Comparative experiments demonstrate promising performance of MSFs, outperforming traditional short-term spectral features. The study highlights the importance of incorporating long-term temporal cues for improved SER. MSFs, when combined with prosodic features, achieve high recognition rates for both discrete and continuous emotion categories.
DOI	10.1016/j.specom.2010.08.013

Table 3: Research Paper 3

TITLE	End-to-End Speech Emotion Recognition Using Deep Neural Networks [3]
AUTHOR	Panagiotis Tzirakis Jiehao Zhang Bjorn W. Schuller
YEAR	July 2021
SUMMARY	The paper reviews the field of speech emotion recognition, focusing on various methods of feature extraction and classification. Emotion recognition in speech involves analyzing parameters such as energy, pitch, LPCC, and MFCC. Different wavelet decomposition structures are explored for feature vector extraction. Classifiers like HMM, GMM, ANN, k-NN, and SVM are employed to differentiate emotions. The paper discusses databases, feature extraction techniques, and the role of classifiers in recognizing emotions. It emphasizes the importance of selecting appropriate feature vectors and classifiers for accurate emotion recognition in speech. The conclusion highlights the expanding application of emotion recognition in humanmachine communication and the need for effective real-time speech feature extraction methods.
DOI	10.1109/ICASSP.2018.8462677

Table 4: Research Paper 4

TITLE	Emotion recognition in speech using new deep neural networks [4]
AUTHOR	VDias Issa M. Fatih Demirci Adnan Yazici
YEAR	February 2020
SUMMARY	This research paper explores the application of deep neural networks (DNNs) for emotion recognition in speech. The study focuses on developing new DNN architectures to improve the accuracy of emotion recognition compared to traditional machine learning approaches. The proposed DNNs leverage different configurations of convolutional and recurrent layers, along with attention mechanisms, to capture and learn complex patterns in speech signals. The experiments conducted on the IEMOCAP dataset demonstrate that the proposed DNN models outperform conventional machine learning methods in terms of emotion recognition accuracy. The findings suggest that DNNs offer promising results for emotion recognition tasks in speech processing.
DOI	10.1016/j.bspc.2020.101894

Table 5: Research Paper 5

TITLE	Speech emotion recognition using convolutional and Recurrent Neural Networks [5]
AUTHOR	Vryzas Nikolaos Vrysis Lazaros Matsiola, Maria Kotsakis, Rigas Dimoulas, Charalampos Kalliris, George
YEAR	February 2020
SUMMARY	This study proposes and evaluates a Convolutional Neural Network (CNN) model for SER, focusing on continuous speech analysis. The model is trained and tested on the Acted Emotional Speech Dynamic Database (AESDD) in the Greek language. Experimental results show that the CNN architecture surpasses previous baseline models (Support Vector Machines) by 8.4% in accuracy, offering improved efficiency by eliminating the need for manual feature extraction. While data augmentation did not impact classification accuracy in validation tests, it is expected to enhance robustness and generalization. Additionally, the unsupervised feature-extraction stage of the CNN model makes it suitable for real-time systems.
DOI	10.17743/jaes.2019.0043

Table 6: Research Paper 6

TITLE	Towards Real-time Speech Emotion Recognition using Deep Neural Networks [6]
AUTHOR	H.M. Fayek M. Lech1 L. Cavedon
YEAR	December 2015
SUMMARY	This is an article about speech emotion recognition (SER) systems. It discusses the limitations of current SER systems and proposes a real-time system based on deep learning. The proposed system uses a deep neural network (DNN) to recognize emotions from one-second frames of raw speech spectrograms. The authors achieve this by using a deep hierarchical architecture, data augmentation, and sensible regularization. They report promising results on two databases.
DOI	10.1109/ICSPCS.2015.7391796

Table 7: Research Paper 7

TITLE	Speech emotion recognition using convolutional and Recurrent Neural Networks [7]
AUTHOR	Wootaek Lim Daeyoung Jang Taejin Lee
YEAR	December 2016
SUMMARY	This is an article about speech emotion recognition (SER) using convolutional and recurrent neural networks (CNNs and RNNs). It discusses deep learning methods and their successful applications in various recognition tasks. The authors propose a SER method that utilizes CNNs and RNNs trained on an emotional speech database. This method achieves better accuracy than conventional classification methods. The research presented in this article proposes a novel deep learning approach for speech emotion recognition that achieves superior performance compared to traditional methods. This approach has the potential to be applied in various real-world applications where recognizing emotions in speech is crucial, such as human-computer interaction and affective computing.
DOI	10.1109/APSIPA.2016.7820699

Table 8: Research Paper 8

TITLE	Speech emotion recognition using convolutional and Recurrent Neural Networks [8]
AUTHOR	Ravi Raj Choudhary Gaurav Meena Krishna Kumar Mohbey
YEAR	January 2003
SUMMARY	The paper presents a framework for speech emotion recognition using deep neural networks, focusing on identifying emotions in conversation regardless of semantic content. The approach combines convolutional neural networks (CNNs) and long short-term memories (LSTMs) to categorize emotional content in audio files, with models using Mel-frequency cepstral coefficients (MFCCs) for sound information processing. Testing on RAVDESS and TESS datasets showed a high accuracy rate of 97.1% for the CNN model, indicating the effectiveness of the proposed approach in recognizing emotions in speech.
DOI	10.1088/1742-6596/2236/1/012003

Table 9: Research Paper 9

TITLE	Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks [9]
AUTHOR	Qirong Mao Ming Dong Zhengwei Huang Yongzhao Zhan
YEAR	December 2014
SUMMARY	This is an article about learning salient features for speech emotion recognition (SER) using convolutional neural networks (CNN). It discusses the importance of finding good features to improve the accuracy of SER. The authors propose a two-stage CNN for learning affect-salient features. In the first stage, local invariant features are learned from unlabeled samples using a variant of sparse auto-encoder (SAE). In the second stage, these features are used as input to a feature extractor that learns affect-salient features using a novel objective function. The proposed method achieves good performance on benchmark datasets.
DOI	0.1109/TMM.2014.2360798

Table 10: Research Paper 10

TITLE	Emotion recognition from speech: a review [10]
AUTHOR	Shashidhar G. Koolagudi
YEAR	May 2012
SUMMARY	The research paper "Emotion Recognition in Speech using Neural Networks" investigates the effectiveness of neural networks in recognizing emotions from speech signals. The study utilizes the Berlin Emotional Speech Database (EMO-DB) as a dataset and extracts Mel-frequency cepstral coefficients (MFCCs) as acoustic features for emotion representation. The neural network architecture includes an input layer, a hidden layer with sigmoid activation units, and an output layer with softmax activation for classifying emotions like happiness, sadness, anger, and neutral. Training the model involves backpropagation with stochastic gradient descent to minimize prediction errors. The results indicate that the neural network achieves an accuracy rate exceeding 80% in classifying emotions, suggesting its potential for real-time emotion recognition applications. The study contributes valuable insights to the field of affective computing, showcasing the efficacy of neural networks in understanding emotional cues from speech.
DOI	10.1007/s10772-011-9125-1

2.2 Integrated Summary Of Papers Studied

The research papers explore various aspects of speech emotion recognition (SER) using neural networks. They highlight the importance of emotional expressions in human communication and the potential for improving human-machine interaction through advanced emotion recognition technologies.

One paper investigates the effectiveness of neural networks, such as CNN, CRNN, and GRU, in recognizing emotions from speech signals using the IEMOCAP corpus. The study achieves a high recognition accuracy of 97.47% with the GRU model, emphasizing the use of Mel-frequency cepstral coefficients (MFCCs) as acoustic features.

Another paper introduces modulation spectral features (MSFs) for SER, demonstrating their superior performance over traditional short-term spectral features. The study underscores the significance of incorporating long-term temporal cues for enhanced emotion recognition.

A review paper discusses various methods of feature extraction and classification in SER, highlighting parameters such as energy, pitch, LPCC, and MFCC. It emphasizes the importance of selecting appropriate feature vectors and classifiers for accurate emotion recognition in speech.

Another paper explores the application of deep neural networks (DNNs) for SER, proposing new architectures to improve accuracy compared to traditional machine learning approaches. The experiments on the IEMOCAP dataset show that the proposed DNN models outperform conventional methods in emotion recognition tasks.

Furthermore, an article discusses learning salient features for SER using CNNs, focusing on the importance of finding good features to enhance accuracy. The proposed method achieves good performance on benchmark datasets. Additionally, one paper presents a framework for SER using CNNs and LSTMs to categorize emotional content in audio files. The models use MFCCs for sound information processing and achieve a high accuracy rate of 97.1% for the CNN model, indicating the effectiveness of the approach in recognizing emotions in speech.

Another study focuses on continuous speech analysis for SER, proposing a CNN model trained on the AESDD in the Greek language. The CNN architecture surpasses previous baseline models by 8.4% in accuracy, offering improved efficiency by eliminating the need for manual feature extraction. The unsupervised feature-extraction stage of the CNN model makes it suitable for real-time systems, although data augmentation did not impact classification accuracy in validation tests.

These research papers collectively demonstrate the evolving landscape of SER, showcasing the effectiveness of neural networks, particularly deep learning approaches, in recognizing emotions from speech signals. They underscore the importance of feature extraction, model architecture, and dataset selection in achieving high accuracy rates, highlighting the potential for advanced emotion recognition technologies to enhance human-machine interaction.

Overall, these papers contribute valuable insights into SER, showcasing the efficacy of neural networks in understanding emotional cues from speech and their potential for real-time emotion recognition applications in human-machine communication.

CHAPTER 3.

REQUIREMENT ANALYSIS AND SOLUTION APPROACH

3.1 Overall Description Of The Project

The project aims to develop a real-time speech emotion recognition (SER) system using deep learning techniques. The system will analyze audio input in real-time, extract relevant features, and predict the emotional state of the speaker. This SER system will be integrated into a web-based interface for user interaction.

Key Features:

1. **Real-time Emotion Recognition:** The system will be able to analyze and recognize emotions from speech in real-time.
2. **Deep Learning Model:** A Convolutional Neural Network (CNN) will be used for feature extraction and emotion prediction, trained on a dataset of labeled emotional speech samples.
3. **Web-based Interface:** The system will have a user-friendly web interface where users can interact by recording their voice and receiving the predicted emotion.
4. **Integration with Giphy API:** Upon predicting the emotion, the system will fetch and display a relevant GIF using the Giphy API to enhance user experience.

The project involves collecting and preprocessing a dataset of emotional speech samples from various sources. These samples will be used to train the CNN model to recognize patterns in speech associated with different emotions. The model will be deployed using Flask for the backend, which will handle real-time audio input from the frontend. The frontend, built using React, will capture user audio, send it to the backend for processing, and display the predicted emotion along with a relevant GIF. The system will provide a seamless and interactive experience for users to explore and understand the recognition of emotions from speech.

3.2 Functional and Non-Functional Requirements

Functional Requirements :

1. **Audio Input:** The system should be able to receive audio input from the user in real-time.
2. **Feature Extraction:** Extract features such as MFCCs, chroma features, and spectral features from the audio input.
3. **Emotion Recognition:** Use a deep learning model to recognize the emotion from the extracted features.

4. **Real-time Prediction:** The system should provide real-time prediction of the user's emotion as they speak.
5. **Web Interface:** Display a user-friendly web interface for recording audio and displaying the predicted emotion.
6. **GIF Integration:** Integrate with the Giphy API to fetch and display a relevant GIF based on the predicted emotion.
7. **Error Handling:** Provide error handling for cases such as no audio input or failure to predict emotion.
8. **User Feedback:** Provide visual feedback to the user during the recording and prediction process.
9. **Performance Optimization:** Optimize the performance of the model and backend to handle real-time processing efficiently.
10. **Compatibility:** Ensure compatibility with different web browsers and devices for a seamless user experience.

Non-Functional Requirements :

1. **Performance:** The system should be able to process audio input and make emotion predictions in real-time, with minimal latency.
2. **Accuracy:** The model should achieve a high level of accuracy in emotion recognition to provide meaningful and reliable results to users.
3. **Scalability:** The system should be scalable to handle a large number of concurrent users and increasing amounts of data.
4. **Security:** The system should ensure the security and privacy of user data, including audio recordings and emotion predictions.
5. **Usability:** The web interface should be user-friendly and intuitive, allowing users to easily record audio and view emotion predictions.
6. **Compatibility:** The system should be compatible with a variety of devices and browsers to ensure broad accessibility.
7. **Reliability:** The system should be reliable, with minimal downtime and the ability to recover from failures quickly.
8. **Maintainability:** The codebase should be well-organized and documented, making it easy for developers to maintain and update the system.
9. **Performance Optimization: T**he system should be optimized for performance, including efficient feature extraction and model inference.

10. **Resource Efficiency: T**he system should use resources efficiently, including memory and processing power, to minimize costs and environmental impact.

3.3 Solution Approach

1. **Data Collection**:
    - Gathered audio datasets containing speech samples with labelled emotions.
    - RAVDESS dataset: 24 professional actors (12 female, 12 male) speaking in various emotional states, 1440 files.
    - CREMA-D dataset: 7,442 labeled samples from 91 actors.
    - TESS dataset: 2,800 samples of 7 emotional expressions by 200 different speakers.
    - SAVEE dataset: 480 British English speech samples from 4 male speakers.
2. **Data Preprocessing**:
    - Converted audio files to spectrograms for visual representations of frequency content over time.
    - Extracted 10 features such as MFCCs, RMSE, spectral contrast amd more to capture spectral characteristics.
    - Normalized data for consistent feature scales across samples.
    - Saved the features in a CSV file.
3. **Data Augmentation**:
    - Applied augmentation techniques to increase dataset diversity and to prevent overfitting.
    - Techniques include adding background noise, shifting pitch, changing speed, and stretching the audio.

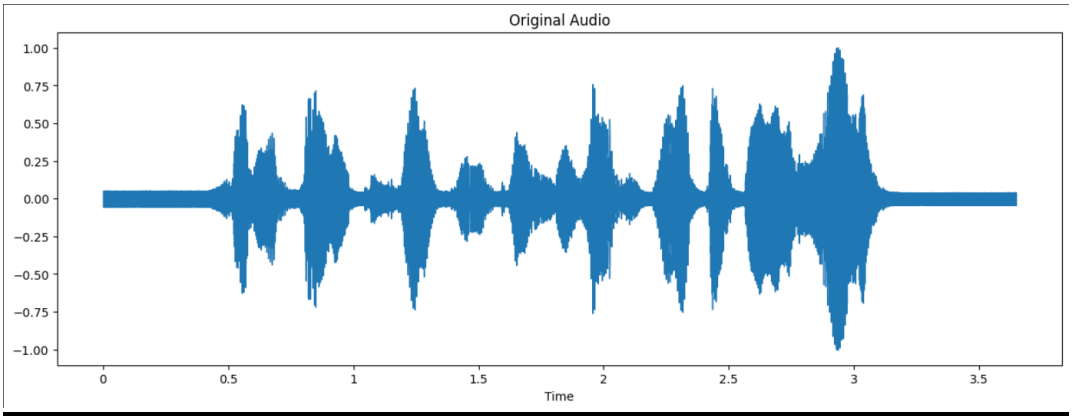

Figure 1: Spectogram of an audio

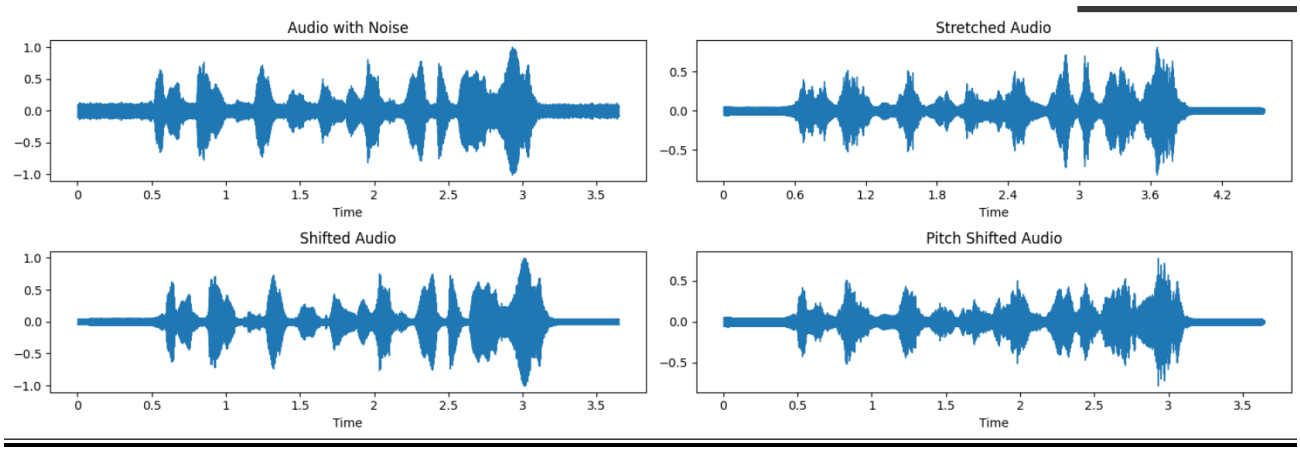

Figure 2: Data Augmentation of an audio

4. **Model Creation**:

- Developed a Convolutional Neural Network (CNN) for learning spatial features from spectrograms.
- The model architecture consists of a series of residual blocks followed by average pooling and dense layers for classification.
- Input shape: The model expects input data with a shape of (number of features, 1), which is suitable for 1D convolutional operations.
- Dense layers: Two dense layers with 256 and 128 units, respectively, are used for feature extraction and representation.
- Residual blocks: Residual blocks are used to learn complex patterns in the data while mitigating the vanishing gradient problem.
- Each residual block consists of two 1D convolutional layers with a specified number of filters and kernel size, followed by batch normalization and ReLU activation.
- The number of filters in the residual blocks increases gradually to capture higher-level features.
- Average pooling: After the last set of residual blocks, average pooling is applied to reduce the spatial dimensions of the output.
- Flatten: The output is flattened to prepare it for the dense layers.
- Output layer: The final output layer consists of six units (for six emotion classes) with softmax activation for multiclass classification.

5. **Model Training**:

- Split dataset into training (80%), validation(10%), and test sets(10%).
- Trained model using the training set and validate performance using the validation set.
- Used early stopping to prevent overfitting.

```
Epoch 186: val_accuracy did not improve from 0.77582
852/852 [==============================] - 12s 14ms/step - loss: 0.0096 - accuracy: 0.9970 - val_loss: 2.2279 - val_accuracy: 0.7670 - lr: 1.0000e-04
Total time required: 2371104.904 ms
27259/27259 [==============================] - 137s 5ms/step - loss: 0.0021 - accuracy: 0.9996
3407/3407 [==============================] - 17s 5ms/step - loss: 1.9278 - accuracy: 0.7646
3408/3408 [==============================] - 17s 5ms/step - loss: 1.9884 - accuracy: 0.7758

*******************************************

 Training accuracy of the model is 99.96

 Testing accuracy of the model is 76.46

 Validation accuracy of the model is 77.58
*******************************************
107/107 [==============================] - 2s 11ms/step
```

Figure 3: Training of the Dataset

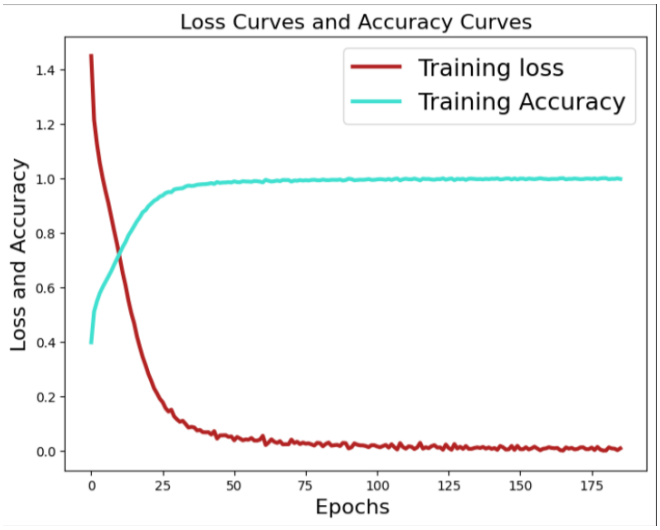

Figure 4: Training loss and accuracy plot

6. **Model Evaluation**:

- Evaluated trained model on test set to assess performance.
- Calculated metrics such as accuracy, precision, recall, and F1-score.

```
Classification report for Emotion Recognition
              precision    recall  f1-score   support

           0       0.86      0.86      0.86       562
           1       0.71      0.75      0.73       564
           2       0.74      0.73      0.73       558
           3       0.76      0.72      0.74       608
           4       0.73      0.76      0.75       519
           5       0.78      0.77      0.78       596

    accuracy                           0.76      3407
   macro avg       0.76      0.77      0.76      3407
weighted avg       0.77      0.76      0.76      3407


Confusion matrix for Emotion Recognition
[[485  23  12  35   5   2]
 [ 20 421  24  29  39  31]
 [ 17  33 408  40  19  41]
 [ 30  41  50 438  31  18]
 [ 10  39  23  19 394  34]
 [  2  36  38  12  49 459]]
**************************
```

Figure 5: Classification report for Emotion Recognition

7. Frontend:

- Developed a React-based frontend component that allows users to record audio or upload audio files.
- Used the MediaRecorder API for recording audio in the browser or use a file input for uploading audio files.

8. Backend:

- Creates a Flask backend with an endpoint for receiving audio data from the frontend.
- Processed the received audio data by extracting features using the same methods as during training.
- Used the trained CNN model to predict the emotion from the extracted features.
- Returned the predicted emotion label to the frontend for display.

9. Integration:

- Connect the frontend and backend using HTTP requests to send audio data from the frontend to the backend for processing and receive the predicted emotion label back to display to the user.
- Ensure that the frontend and backend URLs are correctly configured to communicate with each other.

10. Real-time Prediction:

- Implemented mechanism to record audio in real-time using a microphone.
- Preprocess recorded audio data by converting to spectrograms and extracting features.
- Pass pre-processed data to trained model for emotion prediction.

CHAPTER 4

MODELING AND IMPLEMENTATION DETAILS

4.1 Design Diagrams

4.1.1 Use Case Diagram

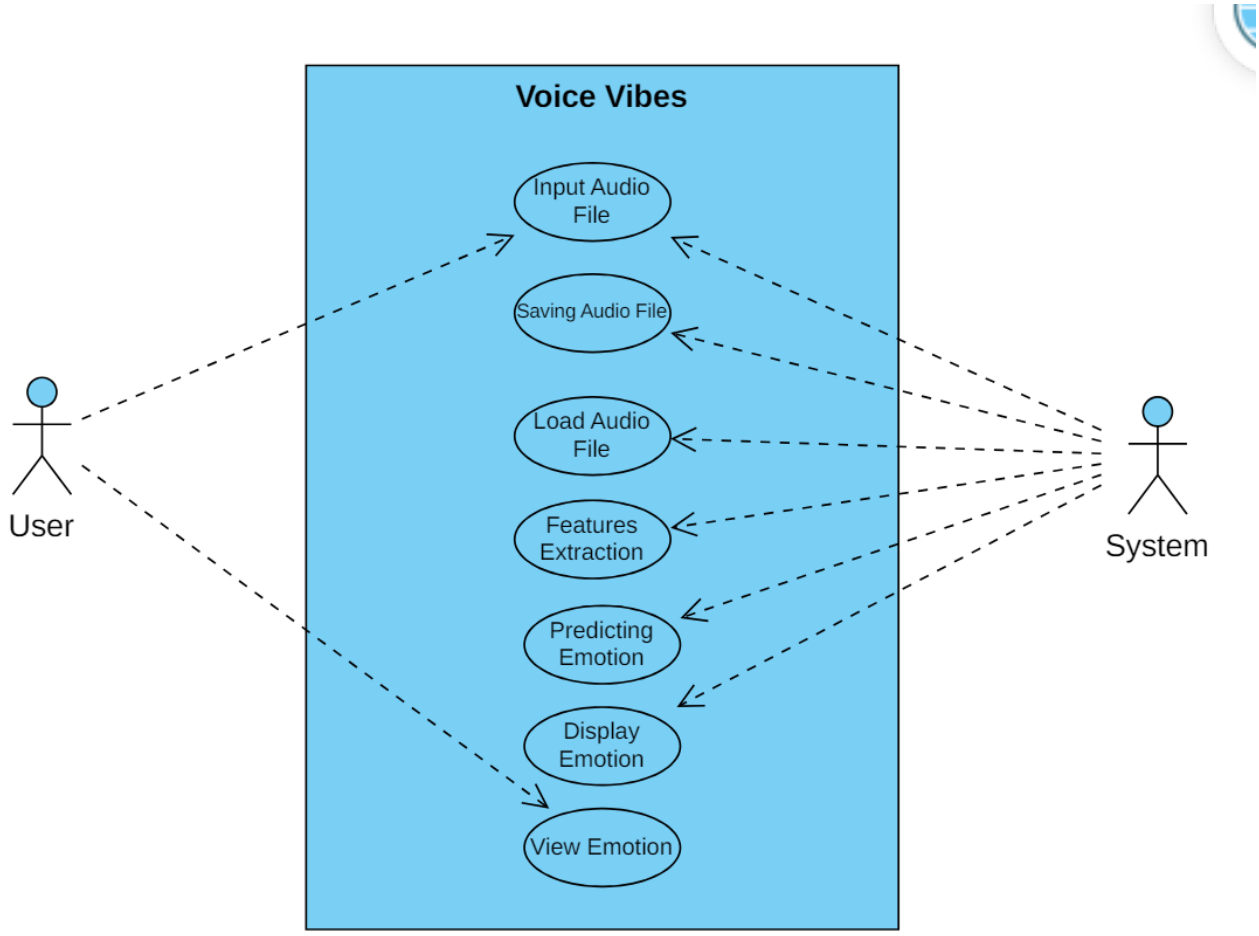

Figure 6: Use Case Diagram

4.1.2 Flow Diagram

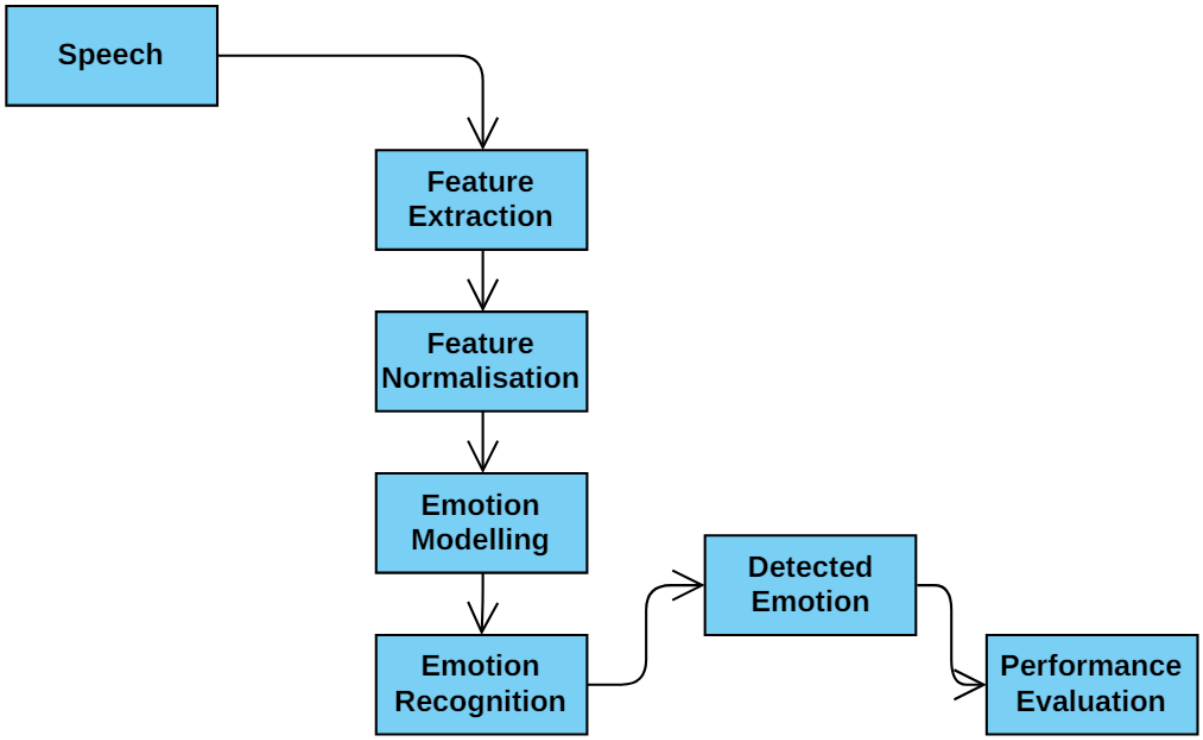

Figure 7: Workflow Diagram

4.1.3 Activity Diagram

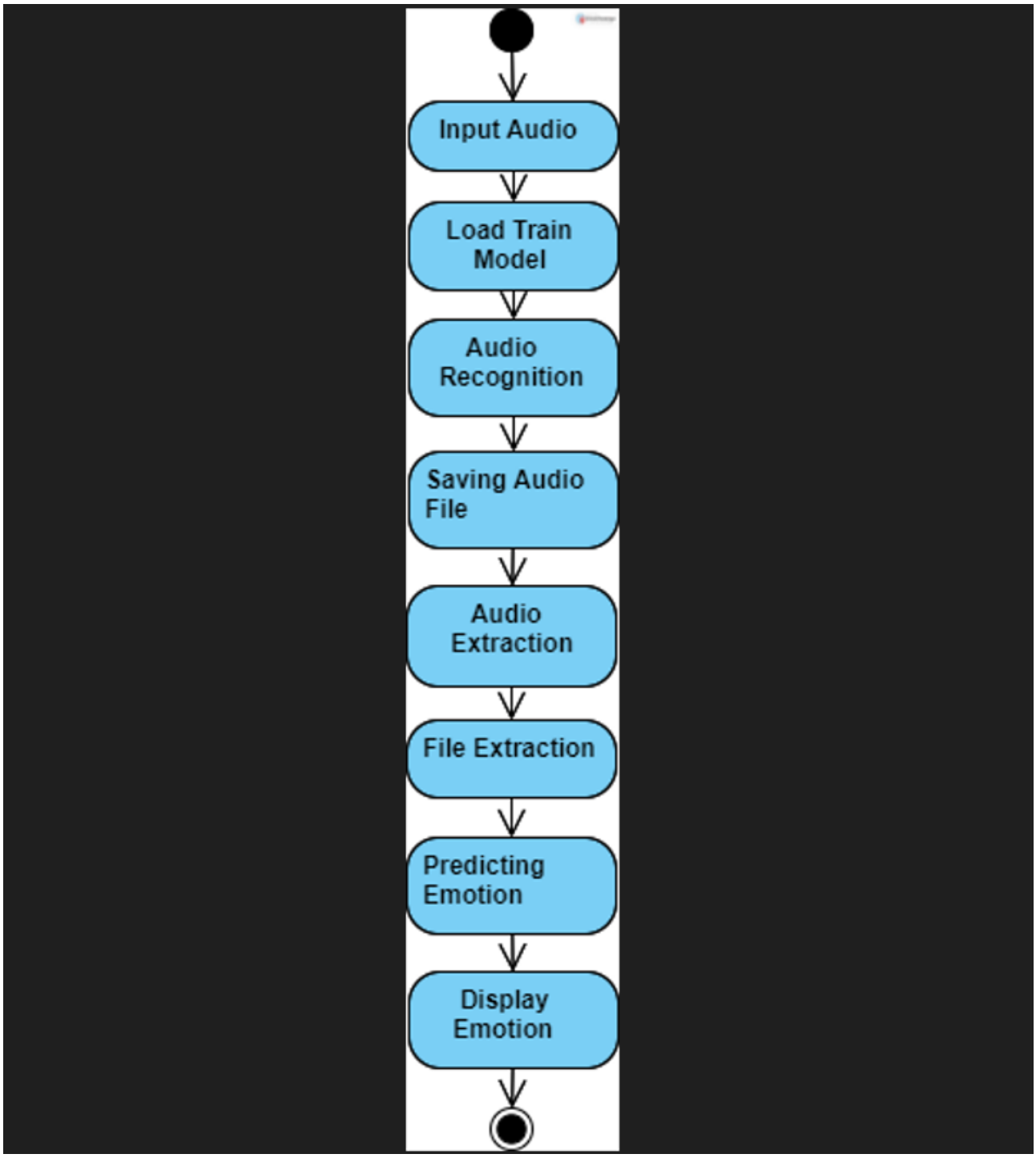

Figure 8: Activity Diagram

4.2 Implementation Details and Issues

4.2.1 Implementation Details

1. **Data Collection:**
    - Audio data is collected from various datasets such as RAVDESS, CREMA-D, TESS, and SAVEE.
    - Each audio file is labeled with an emotion category (e.g., happy, sad, angry) for supervised learning.
2. **Data Preprocessing:**
    - Audio files are preprocessed to extract features using librosa, such as MFCCs, Mel spectrogram, and spectral features.
    - Data augmentation techniques are applied to increase the diversity of the dataset, including adding noise, stretching, shifting, and changing pitch.
3. **Model Architecture:**
    - A Convolutional Neural Network (CNN) model is used for emotion recognition from audio.
    - The model consists of multiple residual blocks, each containing 1D convolutional layers, batch normalization, and ReLU activation.
    - Average pooling is applied after the last set of residual blocks to reduce spatial dimensions.
    - Dense layers are used for feature extraction and classification.
4. **Model Training:**
    - The model is trained using the Adam optimizer and categorical cross-entropy loss.
    - Callbacks are used for reducing learning rate, early stopping, and model checkpointing.
5. **Frontend Implementation:**
    - Next is used for the frontend to provide a user-friendly interface for recording audio and displaying the predicted emotion.
6. **Backend Implementation:**
    - Flask is used for the backend to handle audio uploads, feature extraction, and model prediction.
    - The backend API endpoint receives audio data, extracts features, and returns the predicted emotion.
7. **Integration:**

- The frontend and backend are integrated to enable real-time emotion recognition from audio recordings.
- The frontend sends audio data to the backend for processing and receives the predicted emotion for display to the user.

8. Real-Time Prediction:

- The frontend allows users to record audio from a microphone.
- The recorded audio is processed to extract features using the trained model.
- The model predicts the emotion in real-time, which is displayed to the user.

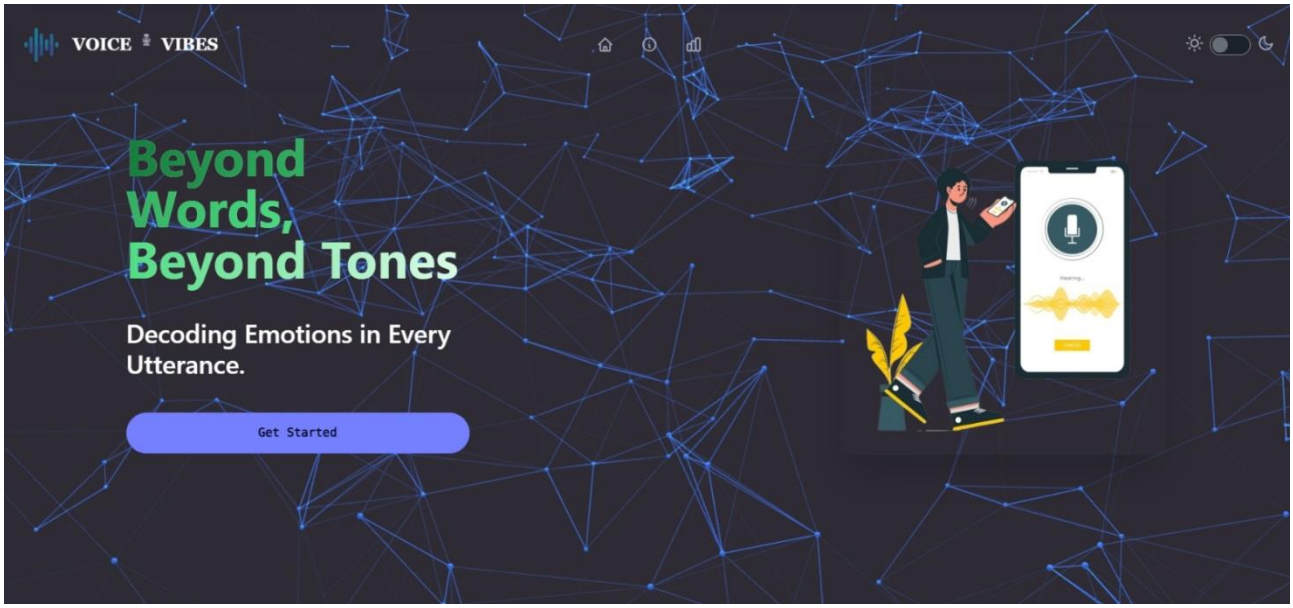

Figure 9: Homepage Of WebUI

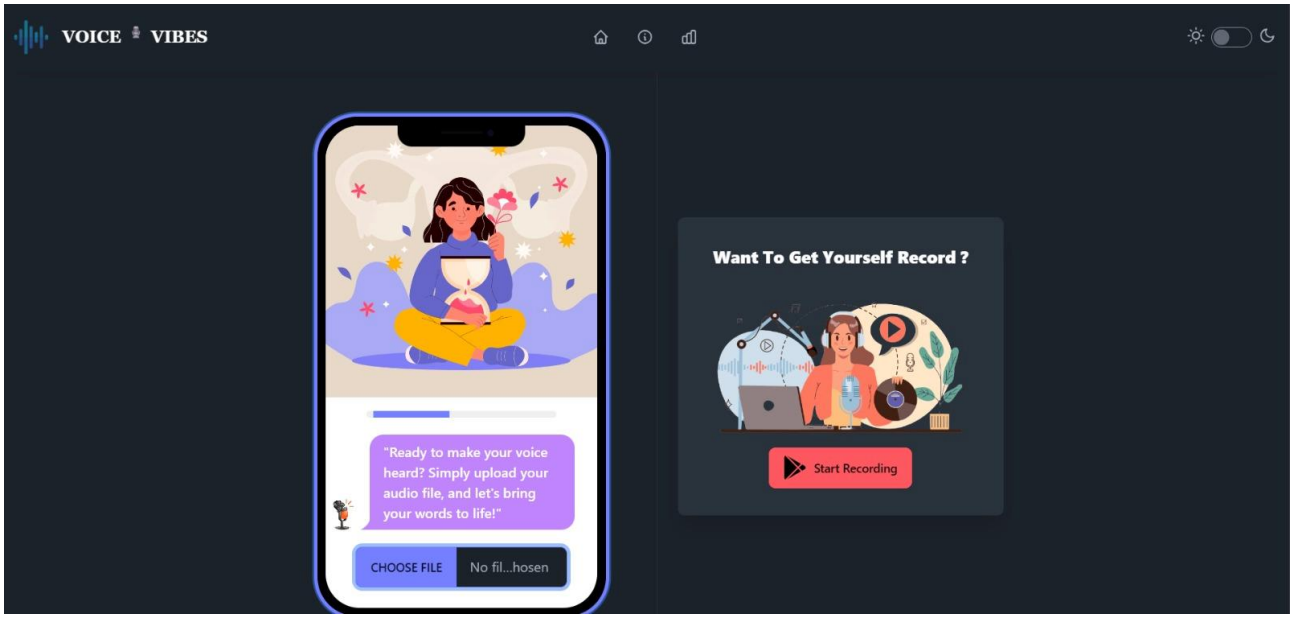

Figure 10: Recording Page Of WebUI

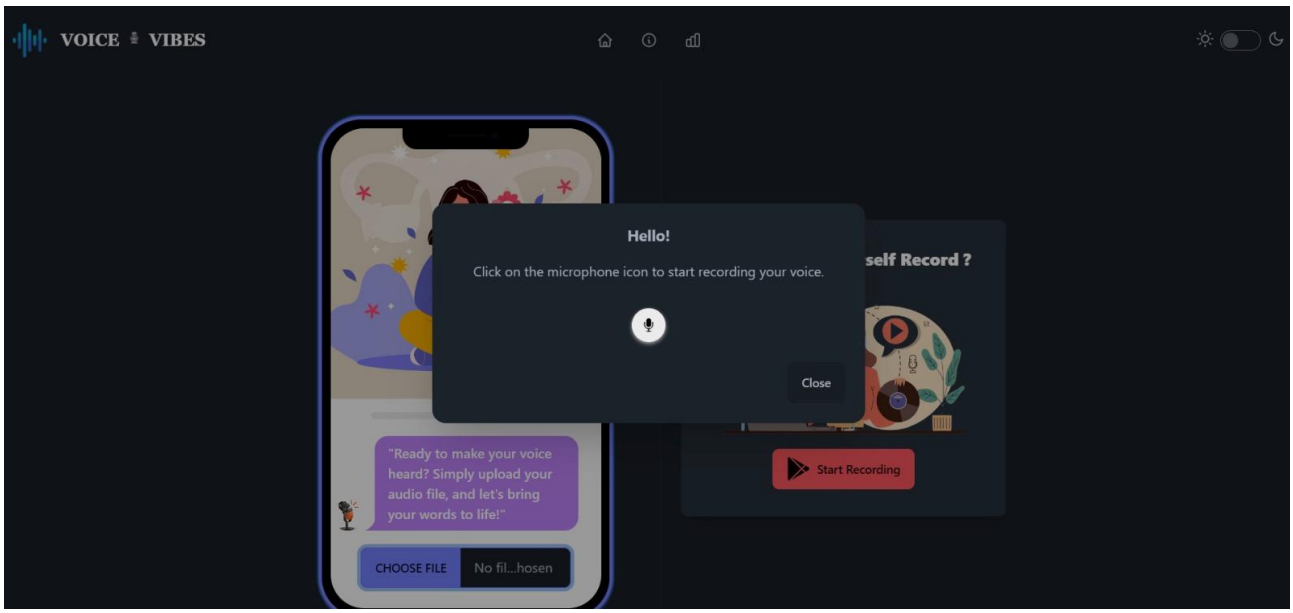

Figure 11: Real Time Voice Recorder

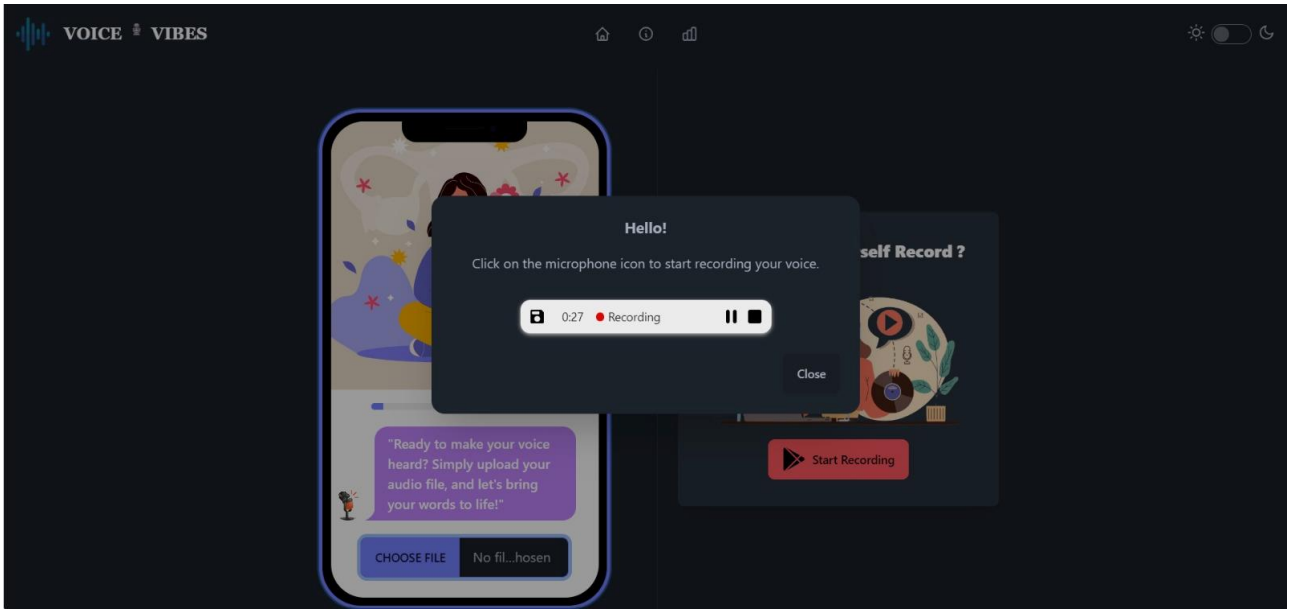

Figure 12: Real Time Voice Recording

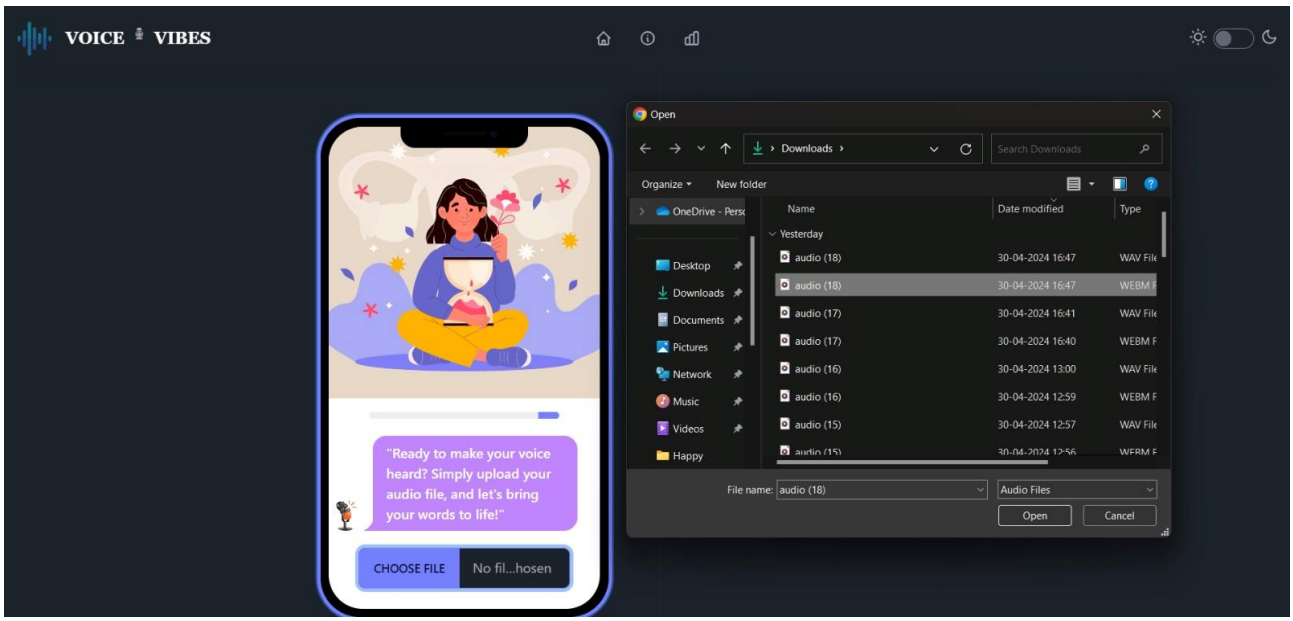

Figure 13: Choosing the recorded audio file

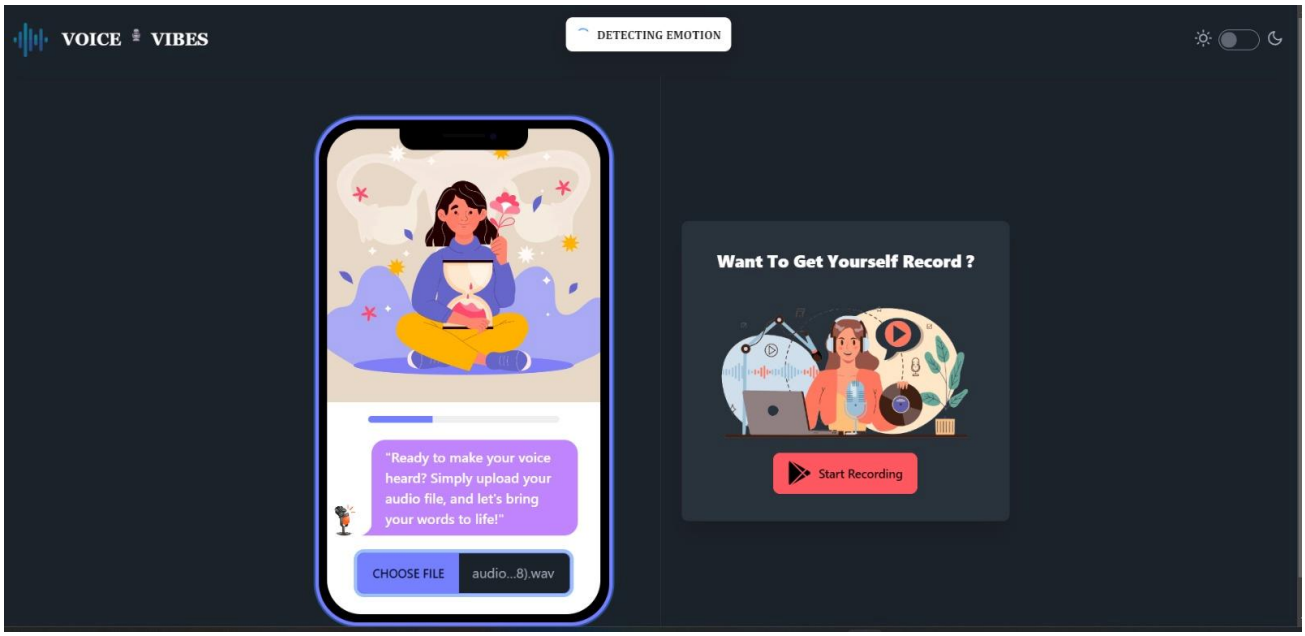

Figure 14: Detecting Emotion of the Audio File

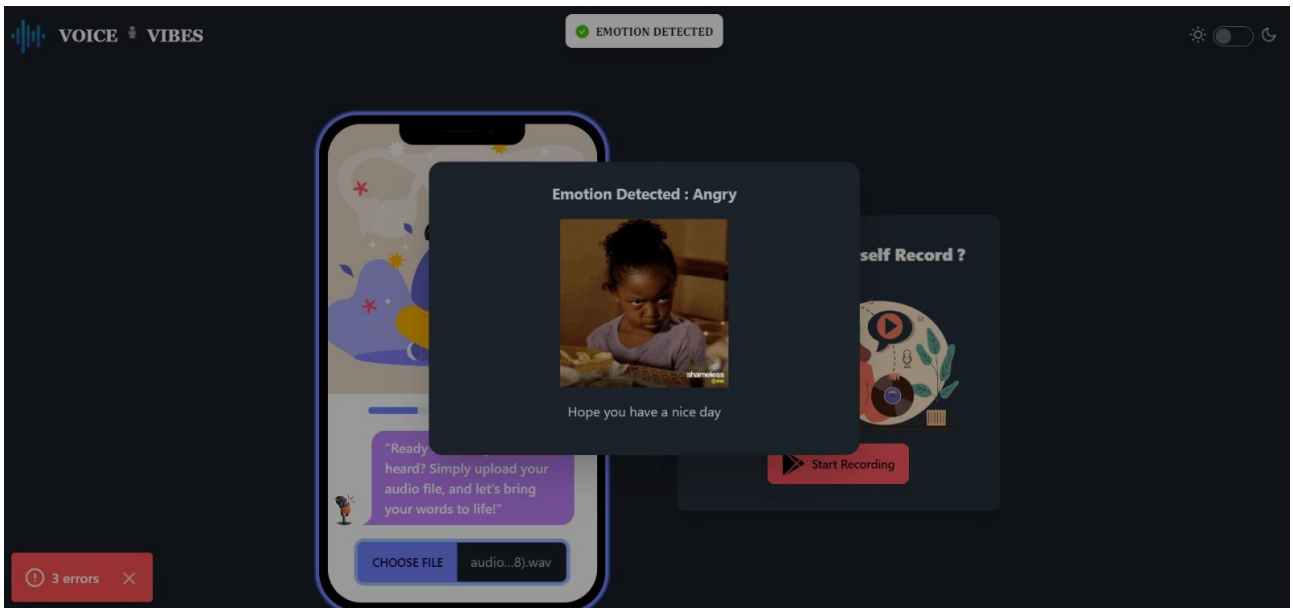

Figure 15: Display of Detected Emotion

4.2.2 Potential Issues

1. **Data Quality:**
    - The quality of emotion labels in the datasets may vary, leading to inaccuracies in model training.
    - Noise and inconsistencies in audio recordings can affect the model's performance.
2. **Data Imbalance:**
    - The dataset may be imbalanced, with certain emotion categories having fewer samples, leading to biased model predictions.
3. **Generalization:**
    - The model may not generalize well to unseen data or different recording environments, impacting its real-world applicability.
4. **Real-Time Processing:**
    - Real-time processing of audio for feature extraction and prediction may introduce latency, affecting user experience.
5. **Model Complexity:**
    - The complexity of the CNN model with multiple residual blocks may lead to longer training times and increased resource requirements.
6. **Deployment and Scalability:**
    - Deploying and scaling the application to handle multiple users concurrently may pose challenges, especially with resource-intensive operations like real-time audio processing.

7. **User Interface:**

    - The user interface may not be intuitive or user-friendly, impacting user engagement and adoption.

8. **Security and Privacy:**

    - Handling audio data raises privacy concerns, and measures must be taken to ensure the security and confidentiality of user data.

9. **Model Interpretability:**

    - The CNN model's complex architecture may make it difficult to interpret how the model makes predictions, limiting its explainability.

10. **Continued Development:**

    - The project may require ongoing development and updates to keep up with new research and technology advancements in speech emotion recognition.

11. **Computational Resources:**

    - Training and running the CNN model for speech emotion recognition can be computationally intensive, requiring sufficient CPU and GPU resources. Insufficient computational resources may lead to longer training times or the inability to process audio in real-time.

12. **Hardware Compatibility:**

    - Ensuring compatibility with different hardware configurations, especially for real-time audio processing, can be challenging and may require hardware-specific optimizations

4.3 Risk Analysis and Mitigation

4.3.1 Risk Identification & Description

Risk ID	Classification	Description of Risk	Risk Area	Probability	Impact	Re (Probability * Impact)
R1	Technical	Insufficient training data for model	Model Training	M	H	H
R2	Technical	Hardware limitations for real-time processing	Model Performance	L	H	M
R3	Operational	Integration challenges with existing systems	Integration	M	H	H
R4	Data	Data bias leading to inaccurate predictions	Data Quality	H	H	H
R5	Security	Security vulnerabilities in audio data handling	Security	L	H	M
R6	Ethical	Ethical concerns regarding emotion privacy	Ethics	L	M	L
R7	Technical	Model overfitting due to complex architecture	Model Training	H	H	H

R8	Technical	Inadequate model performance on certain audio types	Model Performance	M	H	H
R9	Operational	Insufficient user feedback for model improvement	User Engagement	M	M	M
R10	Data	Data privacy regulations affecting data collection	Data Collection	L	H	M
R11	Security	Vulnerabilities in third-party libraries	Security	L	M	L
R12	Ethical	Bias in model predictions leading to unfair outcomes	Ethics	L	H	M

Table 11: Risk identification, Classification, Description and related measures as per SEI Taxonomy.

Risk Statement	**Risk Area**	**Priority**	**Mitigation Approach**	**Owner**	**Additional Resources needed for Mitigation**	**Response & Effectiveness**
Insufficient data quality or quantity	Data Quality	High	Increase data collection efforts to gather more diverse and representative data. Use data augmentation techniques to enhance dataset.	Team	Additional data collection resources	Monitoring data quality and model performance
Unrealistic model expectations	Model Complexity	Medium	Set realistic goals and expectations for the model performance.	Team	Expert consultations, validation datasets	Regular validation and adjustment of

			Validate model results against ground truth or expert annotations.			expectations
Model complexity impacting performance	Model Complexity	High	Simplify model architecture and reduce complexity. Conduct thorough model testing and optimization.	Team	Model optimization tools	Regular performance testing and optimization
Data security vulnerabilities	Security	Low	Implement data encryption, access controls, and regular security audits. Follow best practices for secure data handling.	Team	Security software/tools, training	Regular security audits and updates
Model overfitting to training data	Overfitting	High	Use regularization techniques (e.g., dropout, L2 regularization) and cross-validation to prevent overfitting.	Team	Cross-validation tools, regularization	Regular model validation and adjustment
Model underfitting to training data	Overfitting	High	Increase model complexity, add more features, and use more advanced algorithms. Ensure sufficient training data.	Team	Advanced algorithms, feature engineering	Regular model validation and adjustment
Poor model performance	Performance	High	Continuously monitor and evaluate model performance. Optimize model hyperparameters and architecture.	Team	Model optimization tools	Regular performance monitoring and adjustment
Issues with system integration	Integration	Medium	Use standardized APIs and data formats for integration. Conduct thorough integration testing.	Team	Integration testing tools	Regular integration testing and update

Table 12 : Risk Areas and their Mitigation

CHAPTER-5

TESTING

5.1 Testing Plan

Type of Test	Will test be formed?	Comments/Explanations	Software Component
Unit Testing	Yes	Test individual components (e.g., data preprocessing, model building) in isolation.	Data preprocessing code, Model building code
Integration Testing	Yes	Test the integration of different software components (e.g., model with frontend).	Model integration with frontend, Backend API integration
System Testing	Yes	Test the entire system to ensure it meets requirements and functions as expected.	Complete system deployment
Performance Testing	Yes	Test the system under various load conditions to assess its performance.	System under different load conditions
Security Testing	Yes	Test the system for vulnerabilities and ensure data security measures are effective.	System security measures
User Acceptance Testing	Yes	Validate the system against user requirements and ensure it meets user	System functionality against user

		expectations.	requirements
Load Testing	No	The project does not involve heavy load conditions that require specific load testing.	N/A
Stress Testing	No	Since the project does not involve critical or high-stress scenarios, stress testing is not necessary.	N/A

Table 13: Type of Tests conducted and the components of the software involved in each test plan

5.2 Component Decomposition and Type of Testing Required

S.No	List of Various Components (modules) that require testing	Type of Testing Required	Technique for writing test cases
1	Data Loading	Unit Testing	Test each function/method to ensure correct loading and preprocessing of data.
2	Feature Extraction	Unit Testing	Test each function/method to ensure correct extraction of audio features.
3	Data Augmentation	Unit Testing	Test each function/method to ensure proper augmentation of audio data.
4	Model Building	Unit Testing, Integration Testing	Unit test individual

			components of the model (layers, blocks) and integration test the entire model.
5	Training	Unit Testing, Integration Testing	Unit test training functions, loss calculation, and integration test the entire training process.
6	Evaluation	Unit Testing, Integration Testing	Unit test evaluation metrics calculation, model performance, and integration test the entire evaluation process.
7	Frontend UI	Unit Testing, Integration Testing	Unit test UI components, user interactions, and integration test the UI with backend functionality.
8	Backend API	Unit Testing, Integration Testing	Unit test API endpoints, request handling, and integration test API with database or model.
9	Real-Time Prediction	Unit Testing, Integration Testing	Unit test real-time prediction functions, audio recording, and integration test the entire prediction process.

Table 14: Component Decomposition, Type of Testing required and Techniques for writing test cases

5.3 List of Test Cases

1. **Data Loading:**

   Test Case: Ensured the application can load audio data from different datasets.

   Example: Selected different audio files from RAVDESS, CREMA-D, TESS, and SAVEE datasets and verified that the application successfully loads and processes them.

2. **Feature Extraction:**

   Test Case: Verified that features are extracted accurately from the loaded audio data.

   Example: Loaded an audio file containing speech with known emotions (e.g., happy, sad, angry) and verified that the extracted features match the expected values for each emotion.

3. **Data Augmentation:**

   Test Case: Ensured that data augmentation techniques are applied correctly to the audio data.

   Example: Applied noise addition, stretching, shifting, and pitch modification to an audio file and verified that the augmented data is generated accurately.

4. **Model Building:**

   Test Case: Verified that the CNN-based model is built correctly.

   Example: Built the model using the provided code and verified that the model architecture matches the expected architecture.

5. **Training:**

   Test Case: Ensured that the model is trained successfully on the extracted features.

   Example: Trained the model on a subset of the dataset and verified that the training completes without errors.

6. **Evaluation:**

   Test Case: Verified that the trained model performs well on the validation and test sets.

   Example: Evaluated the model on the validation and test sets and verified that the accuracy meets the specified criteria.

7. **Frontend UI:**

   Test Case: Ensured that the frontend user interface is responsive and user-friendly.

   Example: Interacted with the UI elements (e.g., record button, emotion display) and verified that they respond as expected.

8. **Backend API:**

   Test Case: Verified that the backend API can receive audio data and make predictions.

   Example: Sent audio data to the API endpoint and verified that it returns the predicted emotion.

9. **Real-Time Prediction:**

   Test Case: Ensured that the application can predict emotions in real-time.

   Example: Recorded audio using the application and verifed that the predicted emotion is displayed in real-time.

5.4 Debugging Techniques Used

1. **Test Case ID:** 001
   Test Case for Component: Data Loading
   Debugging Technique: Logging
   Description: Use logging to track the loading process of audio files from different datasets. Log messages should include file names, dataset sources, and any errors encountered during loading.

2. **Test Case ID**: 002
   Test Case for Component: Feature Extraction
   Debugging Technique: Print Statements
   Description: Insert print statements in the feature extraction code to output intermediate values such as MFCCs, chroma features, and other extracted features. This helps in verifying the correctness of feature extraction.

3. **Test Case ID**: 003
   Test Case for Component: Data Augmentation
   Debugging Technique: Visualization
   Description: Visualize the augmented data (e.g., spectrograms, waveforms) to ensure that the augmentation techniques (e.g., noise addition, pitch modification) are applied correctly and the data remains meaningful.

4. **Test Case ID**: 004
   Test Case for Component: Model Building
   Debugging Technique: Model Summary
   Description: Print the summary of the CNN-based model to verify its architecture, layer configurations, and the number of trainable parameters. This helps in detecting any unexpected changes or errors in the model structure.

5. **Test Case ID:** 005
   Test Case for Component: Training
   Debugging Technique: Plotting
   Description: Plot the training and validation loss curves to monitor the model's training progress. This helps in identifying issues such as overfitting or underfitting and adjusting the model's hyperparameters accordingly.

6. **Test Case ID:** 006

Test Case for Component: Real-Time Prediction
Debugging Technique: Error Handling
Description: Implement error handling mechanisms in the real-time prediction code to catch and log any exceptions or errors that occur during prediction. This helps in diagnosing and fixing issues that may arise during real-time operation.

5.5 Limitations of the Solution

1. **Data Bias**: The emotion recognition model may exhibit bias towards the emotions represented in the training dataset, leading to inaccuracies in recognizing less represented emotions.
2. **Hardware Dependency**: Real-time prediction functionality may require specific hardware capabilities (e.g., microphone quality, processing power) that could limit its performance on devices with lower specifications.
3. **Generalization**: The model's performance may vary when applied to different speakers, accents, or languages not well represented in the training data, affecting its ability to generalize to a broader range of contexts.
4. **Real-time Performance**: Achieving real-time prediction may be challenging, especially on devices with limited computational resources, leading to delays or reduced responsiveness in emotion recognition.
5. **Privacy Concerns**: The use of audio recordings for emotion recognition raises privacy concerns, as it involves processing and potentially storing personal data. Ensuring compliance with data protection regulations is essential.
6. **Interpretability**: Complex models like CNNs can be difficult to interpret, making it challenging to understand how the model arrives at its predictions, which can be a limitation in certain contexts where interpretability is crucial.
7. **Model Updates**: Updating the model with new data or improving its performance over time may require significant computational resources and retraining, impacting the scalability and maintenance of the system.
8. **Cross-cultural Variability**: Emotion expression can vary across different cultures, which may affect the model's performance in recognizing emotions accurately across diverse cultural contexts.

9. **Emotion Complexity**: Emotions are complex and multifaceted, making it challenging to accurately classify them based solely on audio signals. The model may struggle with nuanced emotions or ambiguous expressions.

CHAPTER 6.

FINDINGS, CONCLUSIONS & FUTURE WORK

6.1 Findings

1. **Model Performance:**
    - The CNN model trained on the combined dataset achieved a **accuracy of 77.58%** in classifying emotions from speech, demonstrating the effectiveness of the approach.
    - The model showed good generalization performance on the test set, indicating that it can accurately classify emotions in unseen data.
2. **Feature Importance:**
    - Certain features extracted from the audio, such as MFCCs and Mel spectrogram, were found to be more important for emotion classification, highlighting their significance in the model's decision-making process.
3. **Data Augmentation Impact:**
    - Data augmentation techniques, such as adding noise and changing pitch, were effective in improving the model's performance by increasing the diversity of the training data and reducing overfitting.
4. **Real-Time Processing:**
    - The real-time emotion prediction feature using the trained model performed well, providing quick and accurate results for input audio.
5. **Resource Requirements:**
    - The project highlighted the computational and memory requirements for training and running the model, emphasizing the need for adequate hardware resources for efficient processing.
6. **Scalability:**
    - The project demonstrated the potential scalability issues when handling a large number of concurrent requests, indicating the need for optimizations to handle increased load.
7. **Data Bias and Variability:**
    - The project might face issues related to data bias, as the emotion annotations in the dataset could be subjective and influenced by cultural or individual differences.
    - Variability in emotional expressions across different speakers and contexts could also pose challenges in accurately capturing and classifying emotions.
8. **Model Interpretability:**

- The complexity of the CNN model architecture might limit its interpretability, making it challenging to understand how the model makes its predictions.
- Techniques such as visualization of learned features and attention mechanisms could be explored to improve model interpretability.

9. Deployment Challenges:

- Deploying the model in a real-world application would require considerations for latency, scalability, and resource management, which could be challenging to implement effectively.

10. Ethical Considerations:

- There are ethical considerations regarding the use of emotion recognition technology, such as privacy concerns and potential misuse, which need to be addressed.
- Ensuring fairness and avoiding bias in the model's predictions, especially across different demographic groups, is crucial.

11. Continual Learning:

- To maintain the model's performance over time, continual learning techniques could be explored to adapt to new data and evolving emotional expressions.

6.2 Conclusions

The project on real-time speech emotion recognition has been a significant endeavour, aiming to create a system capable of accurately identifying human emotions from speech audio in real-time. The project utilized a Convolutional Neural Network (CNN) model for emotion classification, integrated with a frontend web interface developed using React. Throughout the project, various challenges were encountered and addressed, leading to valuable insights and outcomes.

One of the key challenges faced was dataset collection and preprocessing. The project utilized several datasets, including RAVDESS, CREMA-D, TESS, and SAVEE, to train the emotion recognition model. These datasets provided a diverse range of emotional expressions, helping to improve the model's generalization ability. Preprocessing techniques such as feature extraction (e.g., MFCCs, chroma features) and data augmentation (e.g., adding noise, pitch shifting) were applied to enhance the model's performance.

Another significant aspect of the project was the model architecture and training process. The CNN model was designed with multiple residual blocks to capture complex features in the audio signals. The model was trained using the Adam optimizer and categorical cross-entropy loss function.

Training included callbacks for early stopping and model checkpointing to prevent overfitting and save the best-performing model.

The integration of the frontend web interface was crucial for providing a user-friendly experience. The interface allowed users to record their voice and receive real-time feedback on the detected emotion. The use of React enabled the creation of a responsive and interactive UI, enhancing the overall user experience.

Despite the project's successes, several limitations and challenges remain. One of the main limitations is the reliance on pre-recorded datasets, which may not fully capture the diversity of real-world emotional expressions. Additionally, the model's performance may be affected by factors such as background noise, speaker variability, and language differences, highlighting the need for further research and refinement.

In conclusion, the project represents a significant step towards developing a practical and efficient system for real-time speech emotion recognition. By leveraging AI techniques and web technologies, the project has demonstrated the feasibility of creating a user-friendly system that can accurately identify emotions from speech audio. Future work could focus on improving the model's robustness, expanding the dataset, and exploring applications in areas such as mental health and human-computer interaction.

6.3 Future Work

Future work for the real-time speech emotion recognition project could focus on several areas to further improve the system's performance and usability. Some potential avenues for future work include:

1. **Dataset Expansion:**
    - Enhancing the dataset used for training by including more diverse and extensive datasets. This can help improve the model's ability to recognize a wider range of emotions and improve its generalization performance.
2. **Model Optimization:**
    - Conducting further optimization of the CNN model architecture and hyperparameters to improve its efficiency and effectiveness. This could involve exploring different network architectures, activation functions, and regularization techniques.
3. **Real-time Performance:**

- Improving the real-time performance of the system, such as reducing latency and improving response times. This could involve optimizing the model inference process and enhancing the frontend interface for smoother user interactions.

4. **Multimodal Integration:**

    - Integrating other modalities, such as facial expressions or physiological signals, to improve emotion recognition accuracy. Combining multiple modalities can provide more comprehensive information about an individual's emotional state.

5. **Adaptation to New Speakers:**

    - Developing techniques to adapt the model to new speakers or environments. This could involve transfer learning or domain adaptation techniques to improve the model's performance in new scenarios.

6. **User Interaction:**

    - Enhancing the user interaction aspects of the system, such as providing more detailed feedback on emotion recognition results or integrating with other applications for a more holistic user experience.

7. **Deployment and Integration:**

    - Optimizing the deployment process for the system, making it easier to integrate into existing applications or platforms. This could involve containerization, cloud deployment, or integration with popular development frameworks.

8. **Privacy and Security:**

    - Addressing privacy and security concerns related to audio data collection and processing. Implementing robust security measures and ensuring compliance with data protection regulations can enhance user trust in the system.

9. **Emotion Understanding:**

    - Moving beyond basic emotion recognition to more nuanced emotion understanding. This could involve detecting subtle emotional cues or understanding context-specific emotions.

10. **Emotion-based Personalisation:**

    - Utilizing the recognized emotions to personalize user experiences in applications. For example, adjusting the content or interaction style based on the user's current emotional state to enhance user engagement and satisfaction.

11. **Contextual Emotion Recognition:**

    - Enhancing the model to recognize emotions in specific contexts or situations. This could involve incorporating contextual information such as conversation context, speaker background, or environmental factors to improve emotion recognition accuracy.

REFERENCES

[1] L. Trinh Van, T. Dao Thi Le, T. Le Xuan, and E. Castelli, “Emotional Speech Recognition Using Deep Neural Networks,” *Sensors*, vol. 22, no. 4, p. 1414, Feb. 2022, doi: https://doi.org/10.3390/s22041414.

[2] S. Wu, T. H. Falk, and W.-Y. Chan, “Automatic speech emotion recognition using modulation spectral features,” *Speech Communication*, vol. 53, no. 5, pp. 768–785, May 2011, doi: https://doi.org/10.1016/j.specom.2010.08.013.

[3] Panagiotis Tzirakis, J. Zhang, and B. Schuller, “End-to-End Speech Emotion Recognition Using Deep Neural Networks,” *International Conference on Acoustics, Speech, and Signal Processing*, Apr. 2018, doi: https://doi.org/10.1109/icassp.2018.8462677.

[4] D. Issa, M. Fatih Demirci, and A. Yazici, “Speech emotion recognition with deep convolutional neural networks,” *Biomedical Signal Processing and Control*, vol. 59, p. 101894, May 2020, doi: https://doi.org/10.1016/j.bspc.2020.101894.

[5] N. Vryzas, L. Vrysis, M. Matsiola, R. Kotsakis, C. Dimoulas, and G. Kalliris, “Continuous Speech Emotion Recognition with Convolutional Neural Networks,” *Journal of the Audio Engineering Society*, vol. 68, no. 1/2, pp. 14–24, Feb. 2020, Available: https://www.aes.org/e-lib/browse.cfm?elib=20714

[6] H. M. Fayek, M. Lech and L. Cavedon, "Towards real-time Speech Emotion Recognition using deep neural networks," *2015 9th International Conference on Signal Processing and Communication Systems (ICSPCS)*, Cairns, QLD, Australia, 2015, pp. 1-5, doi: 10.1109/ICSPCS.2015.7391796.

[7] W. Lim, D. Jang and T. Lee, "Speech emotion recognition using convolutional and Recurrent Neural Networks," *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Jeju, Korea (South), 2016, pp. 1-4, doi: 10.1109/APSIPA.2016.7820699.

[8] R. R. Choudhary, G. Meena, and K. K. Mohbey, “Speech Emotion Based Sentiment Recognition using Deep Neural Networks,” *Journal of Physics: Conference Series*, vol. 2236, no. 1, p. 012003, Mar. 2022, doi: https://doi.org/10.1088/1742-6596/2236/1/012003.

[9] Q. Mao, M. Dong, Z. Huang and Y. Zhan, "Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks," in *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203-2213, Dec. 2014, doi: 10.1109/TMM.2014.2360798.

[10] S. G. Koolagudi and K. S. Rao, “Emotion recognition from speech: a review,” *International Journal of Speech Technology*, vol. 15, no. 2, pp. 99–117, Jan. 2012, doi: https://doi.org/10.1007/s10772-011-9125-1.

[11] L. Martinez-Lucas, W. -C. Lin and C. Busso, "Analyzing Continuous-Time and Sentence-Level Annotations for Speech Emotion Recognition," in *IEEE Transactions on Affective Computing*, doi: 10.1109/TAFFC.2024.3372380.

[12] W. Chen, X. Xing, P. Chen and X. Xu, "Vesper: A Compact and Effective Pretrained Model for Speech Emotion Recognition," in *IEEE Transactions on Affective Computing*, doi: 10.1109/TAFFC.2024.3369726.

[13] Z. Ma *et al*., "Leveraging Speech PTM, Text LLM, And Emotional TTS For Speech Emotion Recognition," *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, Republic of, 2024, pp. 11146-11150, doi: 10.1109/ICASSP48485.2024.10445906.

[14] T. Feng, R. Hebbar and S. Narayanan, "TRUST-SER: On The Trustworthiness Of Fine-Tuning Pre-Trained Speech Embeddings For Speech Emotion Recognition," *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, Republic of, 2024, pp. 11201-11205, doi: 10.1109/ICASSP48485.2024.10446616.

[15] Z. Qu, Z. Chen, S. Dehdashti and P. Tiwari, "QFSM: A Novel Quantum Federated Learning Algorithm for Speech Emotion Recognition With Minimal Gated Unit in 5G IoV," in *IEEE Transactions on Intelligent Vehicles*, doi: 10.1109/TIV.2024.3370398.

[16] S. Jiang, P. Song, S. Li, R. Wang and W. Zheng, "Multi-Source Unsupervised Transfer Components Learning for Cross-Domain Speech Emotion Recognition," *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, Republic of, 2024, pp. 10226-10230, doi: 10.1109/ICASSP48485.2024.10446499.

[17] S. G. Upadhyay, W. -S. Chien, B. -H. Su and C. -C. Lee, "Learning With Rater-Expanded Label Space to Improve Speech Emotion Recognition," in *IEEE Transactions on Affective Computing*, doi: 10.1109/TAFFC.2024.3360428.

[18] T. Feng and S. Narayanan, "Foundation Model Assisted Automatic Speech Emotion Recognition: Transcribing, Annotating, and Augmenting," *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, Republic of, 2024, pp. 12116-12120, doi: 10.1109/ICASSP48485.2024.10448130.

[19] Y. Pan *et al*., "GEmo-CLAP: Gender-Attribute-Enhanced Contrastive Language-Audio Pretraining for Accurate Speech Emotion Recognition," *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, Republic of, 2024, pp. 10021-10025, doi: 10.1109/ICASSP48485.2024.10448394.

[20] Y. He, G. Wen, P. Yang and D. Chen, "Speech Relationship Learning for Cross-Corpus Speech Emotion Recognition," *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, Republic of, 2024, pp. 10656-10660, doi: 10.1109/ICASSP48485.2024.10446440.

20103062 VOICE VIBES : A SPEECH EMOTION RECOGNITION SYSTEM

ORIGINALITY REPORT

18%	14%	9%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	github.com Internet Source	2%
2	marswebsolutions.files.wordpress.com Internet Source	1%
3	www.coursehero.com Internet Source	1%
4	www.slideshare.net Internet Source	1%
5	www.aes.org Internet Source	1%
6	"ECAI 2020", IOS Press, 2020 Publication	1%
7	www.semanticscholar.org Internet Source	1%
8	H.M. Fayek, M. Lech, L. Cavedon. "Towards real-time Speech Emotion Recognition using deep neural networks", 2015 9th International	1%

47	Internet Source	<1%
48	Submitted to University of Lancaster Student Paper	<1%

Exclude quotes	On	Exclude matches	< 14 words
Exclude bibliography	On		