SAMSUNG

AI Course

Capstone Project Final Report

Together for Tomorrow!
Enabling People
Education for Future Generations

Emotion Recognition from Urdu Speech Audio Using Deep Learning

12/12/2024

Neural Pair
Rana Uzair Ahmed
Rumail Karim

Content

1. Introduction

1.1. Background Information

1.2. Motivation and Objective

1.3. Members and Role Assignments

1.4. Schedule and Milestones

2. Project Execution

2.1. Data Acquisition

2.2. Training Methodology

2.3. Workflow

3. Results

3.1. Data Preprocessing

3.2. Exploratory Data Analysis (EDA)

3.3. Modeling

3.4. Testing and Improvements

4. Projected Impact

4.1. Accomplishments and Benefits

4.2. Future Improvements

5. Team Member Review and Comment

6. Instructor Review and Comment

1. Introduction

1.1. Background Information

Emotion recognition is an emerging field in artificial intelligence, bridging the gap between human emotions and computational systems. Speech emotion recognition (SER) focuses on detecting emotions from audio signals, making strides in human-computer interaction (HCI) systems. Despite progress, the development of SER systems for underrepresented languages like Urdu remains limited. Leveraging deep learning, this project aims to create an Urdu-based SER model, contributing to inclusivity in AI technologies.

1.2. Motivation and Objective

The primary objective of this project is to develop a model that classifies four basic emotions—Angry, Happy, Neutral, and Sad—from Urdu speech audio. By extracting meaningful features such as Mel-Frequency Cepstral Coefficients (MFCCs) and Mel Spectrograms, the project explores the application of Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and hybrid CNN-LSTM models.
The motivation stems from the growing demand for emotion-aware applications in mental health monitoring, customer service, and other domains where human emotions play a vital role.

1.3. Members and Role Assignments

Rana Uzair Ahmed
Feature Extraction and EDA: Extract audio features (e.g., MFCCs, Mel Spectrograms, etc.) and analyze their patterns.
Model Training: Train and evaluate the LSTM model.
Presentation

Rumail Karim
Data Augmentation and Preparation: Apply augmentation techniques, scale data, and prepare it for modeling.
Model Training: Train and evaluate the CNN and CNN-LSTM models.
Documentation

1.4. Schedule and Milestones

Project Timeline:

- **Total Duration:** 2 weeks
- Start Date: 2nd December 2024

Milestones:

1. Idea Presentation: 4th December 2024
2. Action Plan Submission: 5th December 2024
3. EDA Presentation: 6th December 2024
4. Completion of Code: 10th December 2024
5. Code Submission: 11th December 2024
6. Report Submission: 12th December 2024
7. Final Presentation: 12th December 2024
8. Completion of Project: 12th December 2024

2. Project Execution

2.1. Data Acquisition

The project utilized the URDU-Dataset, comprising 400 audio recordings labeled with four emotions: Angry, Happy, Neutral, and Sad. The data was sourced from Urdu talk shows, featuring 38 speakers (27 male, 11 female). The dataset was accessed from GitHub[1] and organized into folders corresponding to emotions. The audio files, stored in .wav format, were approximately 21 MB per emotion.

[1] *https://github.com/siddiquelatif/urdu-dataset*

2.2. Training Methodology

The training methodology followed a systematic approach. Initially, audio features such as MFCCs, Mel Spectrograms, Chroma Features, RMS Energy, Zero-Crossing Rate, and Spectral properties were extracted to effectively represent speech signals. Data augmentation techniques, including noise addition, time stretching, pitch shifting, and time shifting, were applied to enhance generalization. Three deep learning models—CNN for spatial feature analysis, LSTM for temporal dependencies, and a hybrid CNN-LSTM—were developed and trained. Finally, the models were evaluated using metrics such as accuracy, confusion matrices, precision, and recall.

2.3. Workflow

The workflow for this project began with organizing the audio files and creating metadata to streamline processing. Feature engineering involved extracting and normalizing features using tools like librosa and scikit-learn. Data augmentation was employed to enhance dataset diversity through synthetic variations. Subsequently, CNN, LSTM, and CNN-LSTM models were developed iteratively and trained on the processed data. The final step involved rigorous testing and validation to assess model performance on unseen data.

3. Results

3.1. Data Preprocessing

The data preprocessing phase was crucial to ensure that the dataset was consistent, clean, and ready for feature extraction and model training. This step involved several key stages:

3.1.1. Audio File Handling and Metadata Creation

The first step in the preprocessing pipeline involved organizing and loading the audio files. The dataset consisted of **400 audio recordings** labeled with four emotions: **Angry**, **Happy**, **Neutral**, and **Sad**. These audio files were organized into separate folders for each emotion.

- **Metadata**:
  A metadata DataFrame was created to store key information for each audio file, including:
  - **File Path**: The location of each audio file.
  - **Emotion Label**: The corresponding emotion for each audio file (Angry, Happy, Neutral, Sad).
  - **Duration**: The duration of the audio clip.
  - **Sampling Rate**: The rate at which audio data was sampled.

This metadata helped in organizing the data and allowed easy access to each audio file during feature extraction.

3.1.2. Audio Feature Extraction

Feature extraction is essential for transforming raw audio data into meaningful representations that can be processed by machine learning models. The following audio features were extracted:

- **MFCC (Mel-Frequency Cepstral Coefficients)**:
  MFCCs were extracted to capture the spectral properties of the speech signal, which are highly correlated with human perception of sound. This feature is particularly useful for speech and emotion recognition.
- **Mel Spectrogram**:
  A Mel spectrogram represents the frequency content of the audio over time, mapped to the Mel scale. This helps capture both frequency and time-based changes in the audio.
- **Zero-Crossing Rate (ZCR)**:
  ZCR measures the number of times the signal crosses the zero amplitude axis. It is often used to determine the noisiness or sharpness of the sound.
- **RMS Energy**:
  The Root Mean Square (RMS) Energy calculates the average energy of the audio signal, which reflects the loudness or intensity of the sound.
- **Spectral Centroid and Bandwidth**:
  Spectral centroid measures the "center of mass" of the spectrum and is often associated with the perceived brightness of sound. Spectral bandwidth represents the range of frequencies around the spectral centroid, which gives an idea of the tonal width.
- **Chroma Features**:
  Chroma features capture the harmonic content of the audio by representing the intensity of the 12 pitch classes (C, C#, D, etc.). These features are particularly useful for detecting musical and tonal patterns in speech.

3.1.3. Data Augmentation

To make the model more generalizable and robust to real-world variations, several **data augmentation techniques** were applied to the audio data:

- **Noise Addition**: Adding slight background noise to the audio helps the model handle real-world conditions where audio is not always clean.
- **Time Stretching**: Speeding up or slowing down the audio without changing its pitch helps the model learn to recognize emotions across varying speech rates.
- **Pitch Shifting**: Altering the pitch of the audio samples helps the model generalize across different vocal tones and accents.

3.1.4. Data Scaling and Reshaping

After feature extraction and augmentation, the data was scaled using **StandardScaler** to normalize the values across all features. This step ensures that all features are on the same scale, which is important for effective model training.

- **Reshaping**: The data was reshaped to be compatible with the models. For **CNN**, the data was reshaped into 2D arrays, and for **LSTM**, it was reshaped into sequences for temporal analysis.

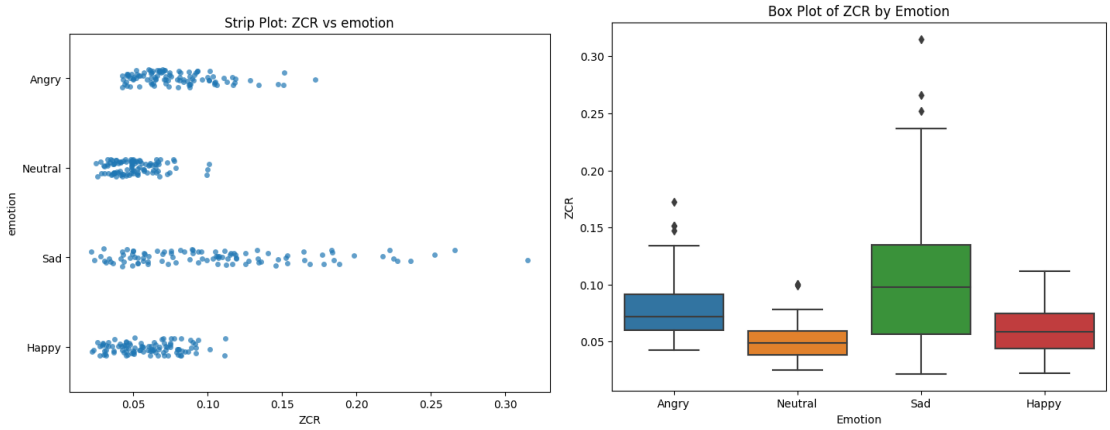
3.2. Exploratory Data Analysis (EDA)

In the EDA phase, the main goal was to analyze the distributions and patterns of the extracted features across different emotions. This involved generating various visualizations to better understand the behavior of the features.

3.2.1. Zero Crossing Rate (ZCR)

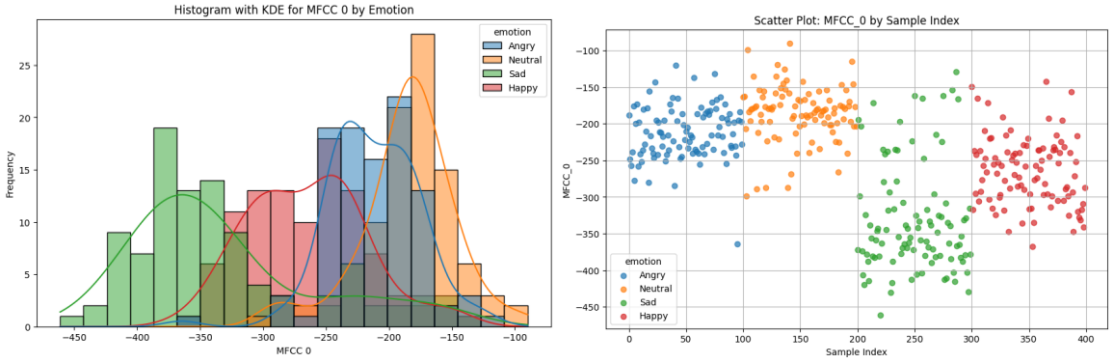
The Zero Crossing Rate (ZCR) measures how often the audio signal crosses zero amplitude. It is typically used to characterize noisiness or sharpness in the signal. The ZCR values were

calculated for each audio sample, and their distribution across emotions was visualized to identify any patterns in the data.



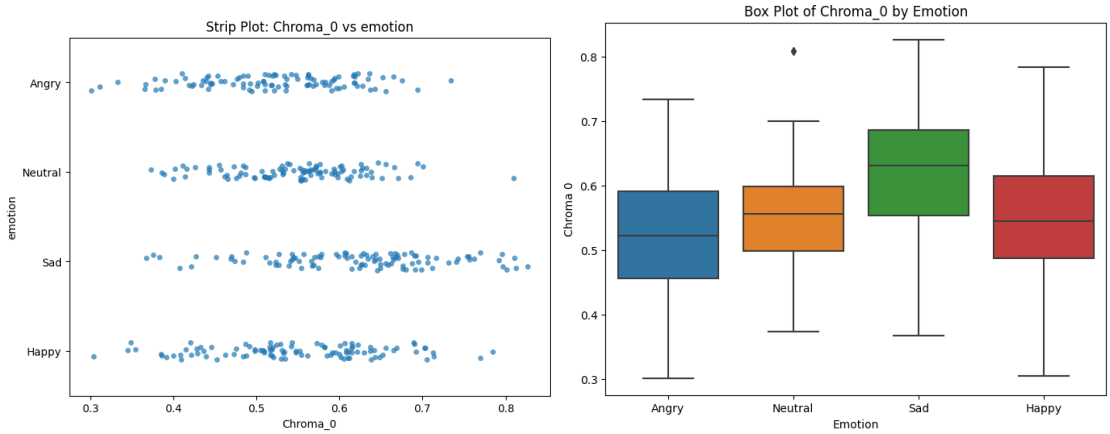
3.2.2. Mel-Frequency Cepstral Coefficients (MFCCs)

MFCCs represent the spectral properties of the audio signal. They are widely used in speech processing tasks because they capture the frequency characteristics in a way that mirrors human auditory perception. The MFCCs were extracted for each audio sample, and specific coefficients (e.g., MFCC_0, MFCC_12) were analyzed to observe their distribution and variation across different emotions.



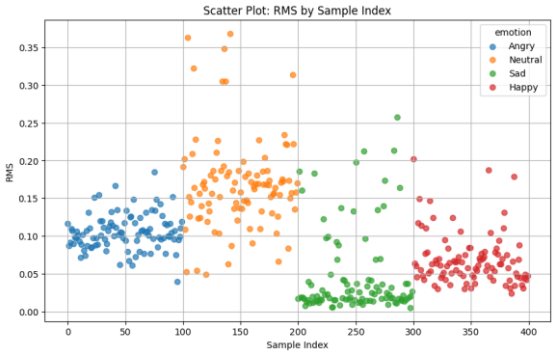
3.2.3. Chroma Features

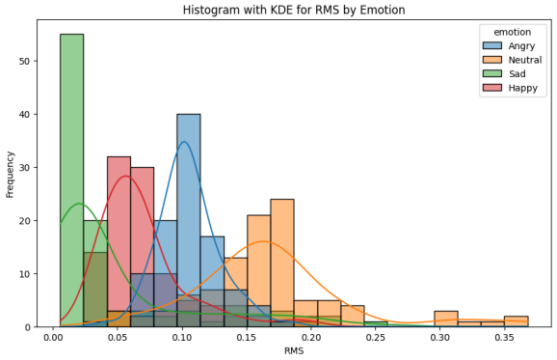
Chroma features capture the harmonic content of the audio by representing the energy in each of the 12 pitch classes (C, C#, D, etc.). These features are useful for identifying musical and tonal qualities. Chroma features were calculated for each sample, and their distribution across emotions was analyzed to check for patterns related to tonal content.



3.2.4. Root Mean Square (RMS) Energy

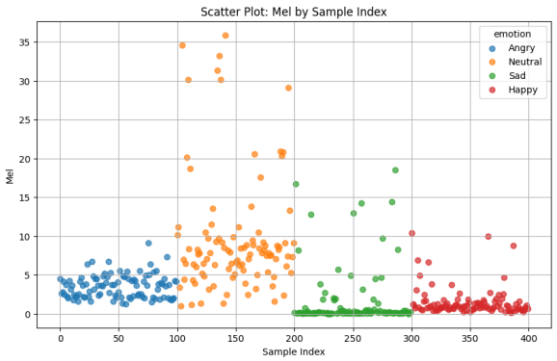
RMS energy measures the loudness or intensity of the audio signal. It is a common feature in speech and emotion recognition tasks, as emotional speech tends to have varying levels of energy. The RMS energy values were calculated for each audio sample, and the distribution of energy levels across emotions was analyzed.





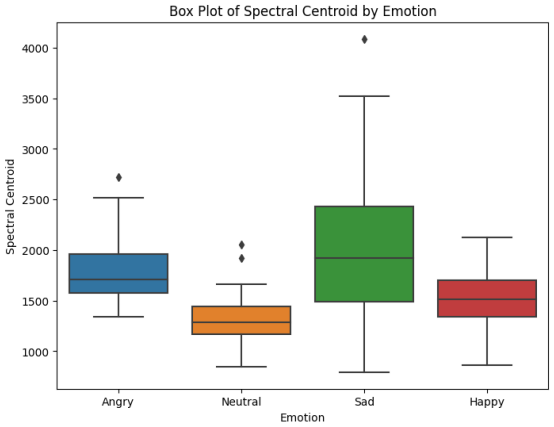
3.2.5. Mel Spectrogram

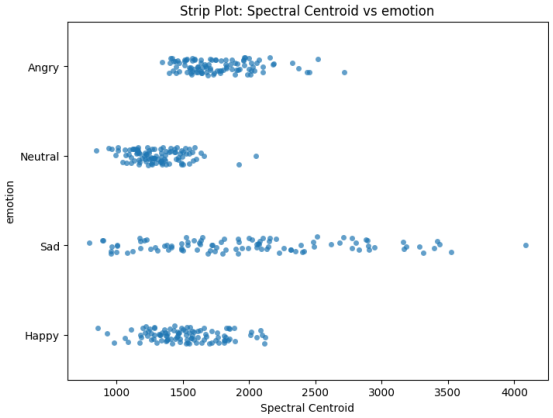
A Mel spectrogram represents the audio's frequency content over time, mapped to the Mel scale. It is widely used in speech analysis due to its alignment with human hearing. The Mel spectrogram was computed for each audio sample, and its time-frequency characteristics were visualized. The focus was on understanding how these spectrograms differ across emotions.



3.2.6. Spectral Centroid

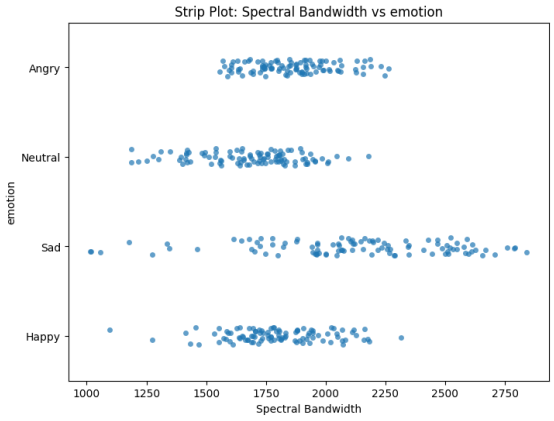
Spectral centroid represents the "center of mass" of the spectrum and is related to the perceived brightness of the sound. A higher spectral centroid is often associated with brighter sounds. The spectral centroid was calculated for each sample, and the distribution of this feature across emotions was analyzed.

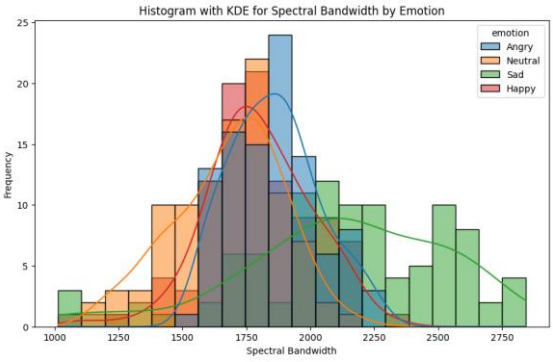




3.2.7. Spectral Bandwidth

Spectral bandwidth measures the width of the frequency distribution and provides insight into the "sharpness" or "dullness" of the sound. Spectral bandwidth was calculated for each audio sample, and its distribution across emotions was visualized.





3.3. Modeling

Three deep learning models were developed to classify emotions from Urdu speech audio. The models used the features extracted during preprocessing and were trained and evaluated systematically:

3.3.1. Convolutional Neural Network (CNN)

The CNN model analyzed spatial patterns in audio features like MFCCs and Mel Spectrograms.

- **Architecture**:
  - Convolutional layers with ReLU activation to detect important patterns in the input features.
  - Max Pooling layers to reduce the size of feature maps and retain key information.
  - Dense layers for emotion classification.
- **Key Features**: The CNN focused on learning spatial features effectively and handled variations in data well.
- **Performance**: The CNN model outperformed the other two models, achieving the highest accuracy among all approaches and providing consistent classification across all emotion categories.

3.3.2. Long Short-Term Memory (LSTM)

The LSTM model was designed to capture temporal dependencies in audio signals, making it well-suited for sequential data.

- **Architecture**:
  - Two LSTM layers stacked to process sequences over time.
  - Dense layers to map the processed sequences to emotion labels.
- **Key Features**: LSTM captured changes in the audio signal over time but was less effective at detecting spatial patterns in the features.
- **Performance**: While LSTM showed good results, it did not match the CNN model in terms of accuracy and consistency.

3.3.3. Hybrid CNN-LSTM Model

The hybrid model combined CNN and LSTM to take advantage of both spatial and temporal feature analysis.

- **Architecture**:
  - CNN layers processed the spatial features from the input.
  - LSTM layers analyzed the temporal relationships in the features extracted by the CNN.
  - Dense layers performed the final classification.
- **Key Features**: This model attempted to combine the strengths of CNN and LSTM to understand the data more comprehensively.
- **Performance**: The hybrid model performed better than LSTM alone but did not surpass the CNN due to its increased complexity.

3.5. Testing and Improvements

After training, the models were thoroughly tested to evaluate their performance on unseen data. The goal was to assess how well each model could classify emotions and identify areas for improvement.

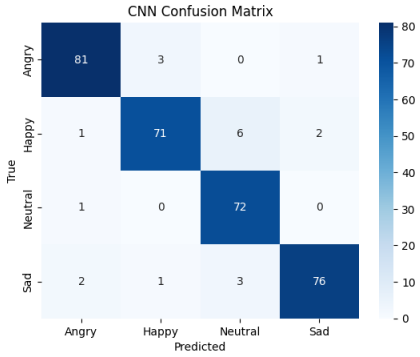
- **Test Set**: The dataset was split into training (80%) and test (20%) sets, ensuring a balanced representation of all emotions.
- **Evaluation Metrics**: The models were assessed using several metrics to measure their classification performance:
  - **Accuracy**: The overall percentage of correct predictions.
  - **Precision**: The ability of the model to correctly identify each emotion.
  - **Recall**: The model's ability to correctly detect all instances of a given emotion.
  - **F1-Score**: The balance between precision and recall.
  - **Confusion Matrix**: A confusion matrix was generated for each model to visualize the true positives, false positives, true negatives, and false negatives for each emotion class.

Results

- **CNN**:
  - Accuracy: **93.75%**
  - The CNN model performed consistently well across all emotion categories, showing a strong ability to detect both distinct and subtle features of each emotion.
  - Classification Report and Confusion Matrix:

	Precision	Recall	F1-Score	Support
Angry	0.95	0.95	0.95	85
Happy	0.95	0.89	0.92	80
Neutral	0.89	0.99	0.94	73
Sad	0.96	0.93	0.94	82

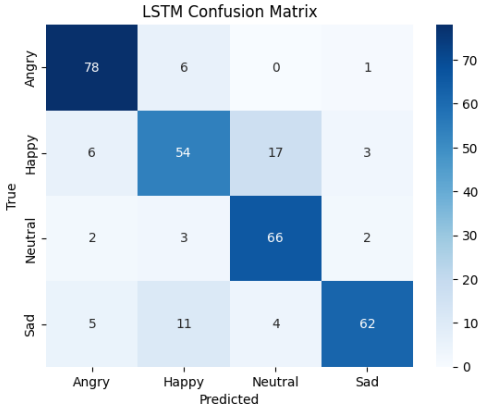
Accuracy			0.94	320
Macro avg	0.94	0.94	0.94	320
Weighted avg	0.94	0.94	0.94	320



- **LSTM**:
  - Accuracy: **81.25%**
  - The LSTM model was effective at capturing temporal dependencies in the audio but struggled with distinguishing between emotions with similar vocal patterns, like Angry vs. Sad.
  - Classification Report and Confusion Matrix:

	Precision	Recall	F1-Score	Support
Angry	0.86	0.92	0.89	85
Happy	0.73	0.68	0.70	80
Neutral	0.76	0.90	0.82	73
Sad	0.91	0.83	0.83	82

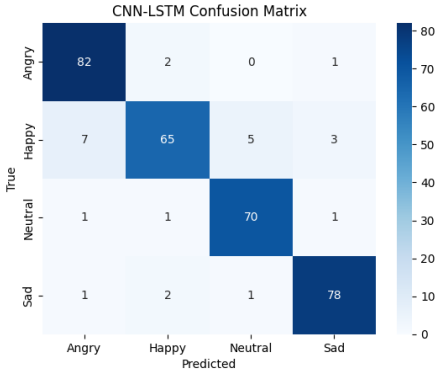
Accuracy			0.81	320
Macro avg	0.81	0.81	0.81	320
Weighted avg	0.82	0.81	0.81	320



- **CNN-LSTM Hybrid**:
  - Accuracy: **92.19%**
  - The hybrid model combined the strengths of both CNN and LSTM, but did not surpass the performance of CNN alone. It took longer to train without providing significant improvements in accuracy.
  - Classification Report and Confusion Matrix

	Precision	Recall	F1-Score	Support
Angry	0.90	0.96	0.93	85
Happy	0.93	0.81	0.87	80
Neutral	0.92	0.96	0.94	73
Sad	0.94	0.95	0.95	82

Accuracy			0.92	320
Macro avg	0.92	0.92	0.92	320
Weighted avg	0.92	0.92	0.92	320



4. Projected Impact

4.1. Accomplishments and Benefits

The project successfully developed deep learning-based emotion recognition models for Urdu speech, achieving promising accuracy. The study highlighted the importance of feature engineering and data augmentation in improving model generalization. The models' potential applications include mental health monitoring, where it could identify emotional states for timely interventions, customer service, enhancing interactions through emotion-aware AI, and promoting language inclusivity by addressing the underrepresentation of Urdu in AI systems.

4.2. Future Improvements

Dataset Expansion: Increasing the dataset size and diversity for improved generalization.
Additional Emotions: Expanding the classification to include a broader range of emotions.

5. Team Member Review and Comment

NAME	REVIEW and COMMENT
Rumail Karim	I gained experience in data preprocessing and model development. This project gave me a deeper understanding of emotion recognition using deep learning.
Rana Uzair Ahmed	This project helped me improve my skills in feature extraction and model training, especially with audio data. It was a valuable learning experience.

6. Instructor Review and Comment

CATEGORY	SCORE	REVIEW and COMMENT
IDEA	/30	
CODING	/20	
RESULT	/10	
PROJECT MANAGEMENT	/20	
PRESENTATION & REPORT	/20	
TOTAL	/100	