SRI RAMACHANDRA
INSTITUTE OF HIGHER EDUCATION AND RESEARCH
(Category - I Deemed to be University) Porur, Chennai
SRI RAMACHANDRA FACULTY OF ENGINEERING AND TECHNOLOGY

SPEECH EMOTION RECOGNITION

CSE 380 NATURAL LANGUAGE PROCESSING

PROJECT REPORT

Submitted by

AKSHAYA KEERTHI P – E0322048

RISHITHA T – E0322026

BHARGOW N – E0322046

In partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

(Artificial Intelligence and Data Analytics)

Sri Ramachandra Faculty of Engineering and Technology

Sri Ramachandra Institute of Higher Education and Research, Porur, Chennai - 600116

JULY ,2025

SRI RAMACHANDRA
INSTITUTE OF HIGHER EDUCATION AND RESEARCH
(Category - I Deemed to be University) Porur, Chennai
SRI RAMACHANDRA FACULTY OF ENGINEERING AND TECHNOLOGY

SPEECH EMOTION RECOGNITION

CSE 380 NATURAL LANGUAGE PROCESSING

PROJECT REPORT

Submitted by

AKSHAYA KEERTHI P – E0322048

RISHITHA T – E0322026

BHARGOW N – E0322046

In partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

(Artificial Intelligence and Data Analytics)

Sri Ramachandra Faculty of Engineering and Technology

Sri Ramachandra Institute of Higher Education and Research, Porur, Chennai -600116

JULY ,2025

SRI RAMACHANDRA
INSTITUTE OF HIGHER EDUCATION AND RESEARCH
(Category - I Deemed to be University) Porur, Chennai
SRI RAMACHANDRA FACULTY OF ENGINEERING AND TECHNOLOGY

BONAFIDE CERTIFICATE

Certified that this project report ”SPEECH EMOTION RECOGNITION” Is the bonafide record of work done by **“BHARGOW N – E0322046, RISHITHA T – E0322026, AKSHAYA KEERTHI P – E0322048”** who carried out the internship work under my supervision.

Signature of the Supervisor
Coordinator

Dr. S. Suja Golden Shiney

Assistant Professor,

Department of Computer Science and Engineering

Sri Ramachandra Faculty of Engineering and Technology,

SRIHER, Porur, Chennai-600 116.

Signature of Programme

Dr. A. Satya

Professor,

Department of Computer Science and Engineering

Sri Ramachandra Faculty of Engineering and Technology,

SRIHER, Porur, Chennai-600 116.

Evaluation Date:

ACKNOWLEDGEMENT

I express my sincere gratitude to our Chancellor, Vice-Chancellor and our sincere gratitude to Our Dean **Prof. T. Ragunathan,** for his support and for providing the required facilities for carrying out this study. I wish to thank my faculty supervisor(s), **Dr. S. Suja Golden Shiny,** Department of Computer Science and Engineering, Sri Ramachandra Faculty of Engineering and Technology for extending help and encouragement throughout the project. Without his/her continuous guidance and persistent help, this project would not have been a success for me.

I am grateful to all the members of Sri Ramachandra Faculty of Engineering and Technology, my beloved parents, and friends for extending the support, who helped us to overcome obstacles in the study.

TABLE OF CONTENTS

ABSTRACT

Speech Emotion Recognition (SER) represents a fascinating computational challenge that I find particularly compelling due to its intersection of human psychology and machine learning. In my work, I've focused on automatically identifying and classifying human emotions from speech signals by leveraging the rich acoustic information embedded in vocal expressions. I believe that emotions manifest distinctly through various acoustic features such as pitch variations, intensity fluctuations, spectral characteristics, and prosodic patterns, which provide a reliable foundation for computational analysis.

For my research, I chose to implement a Random Forest model, which I found to be particularly effective for this classification task. While many researchers gravitate toward deep learning architectures like CNNs, RNNs, and transformer models, I was drawn to Random Forest due to its interpretability, robustness to overfitting, and excellent performance with the feature set I extracted. The ensemble nature of Random Forest allowed me to capture complex relationships between different acoustic features while maintaining computational efficiency.

I utilized the TESS (Toronto Emotional Speech Set) dataset for my experiments, which I selected because of its high-quality recordings and clear emotional labels. Working with TESS gave me valuable insights into how different emotions are acoustically encoded, particularly across the range of emotions it covers. The dataset's controlled recording conditions helped me focus on the core acoustic-emotional relationships without dealing with excessive noise or variability.

INTRODUCTION

Speech Emotion Recognition (SER) is a rapidly evolving field that combines signal processing, machine learning, and affective computing to automatically identify and classify human emotions from speech signals. As human speech naturally conveys emotional information through acoustic cues such as pitch variations, speaking rate, vocal intensity, and spectral characteristics, SER systems aim to decode these paralinguistic features to understand the speaker's emotional state.

The significance of speech emotion recognition extends far beyond academic research, with practical applications spanning human-computer interaction, mental health monitoring, customer service analytics, and educational technology. In an increasingly digital world where voice-based interfaces are becoming ubiquitous, the ability to recognize and respond to human emotions through speech represents a crucial step toward more empathetic and intelligent systems.

Traditional SER approaches relied on handcrafted acoustic features such as Melfrequency Cepstral Coefficients (MFCCs), pitch statistics, energy measures, and formant frequencies, combined with classical machine learning algorithms like Support Vector Machines and Hidden Markov Models. However, the advent of deep learning has revolutionized the field, enabling end-to-end systems that automatically learn relevant features from raw audio using Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and transformer architectures.

Environmental factors further complicate real-world deployment. Systems trained on clean, studio-quality recordings often struggle with noisy, compressed, or reverberant audio typical of telecommunication applications. Cross-lingual and cross-cultural generalization remains challenging, as emotional expression varies significantly across different languages and cultural contexts.

Human-computer interfaces increasingly incorporate emotion recognition to create more natural and responsive interactions. Virtual assistants and chatbots use emotional cues to adapt their communication style, while gaming and entertainment applications create more immersive experiences by responding to player emotional states.

LITERATURE REVIEW

3.1. Early Foundations and Feature-Based Approaches

The foundational work in speech emotion recognition emerged in the 1990s with pioneering studies by Dellaert et al. (1996) and Cowie et al. (2001), who established the theoretical framework for extracting emotional information from acoustic features. Early research focused on handcrafted feature extraction, with Schuller et al. (2003) introducing comprehensive feature sets including prosodic, spectral, and voice quality measures. The widely-adopted openSMILE toolkit by Eyben et al. (2010) standardized feature extraction processes, enabling reproducible research across the community.

3.2. Deep Learning Revolution

The paradigm shift to deep learning began with Stuhlsatz et al. (2011), who first applied deep belief networks to SER, demonstrating superior performance over traditional methods. Huang et al. (2014) introduced Convolutional Neural Networks for spectrogram-based emotion recognition, while Mirsamadi et al. (2017) pioneered the use of attention mechanisms for temporal emotion dynamics modeling.

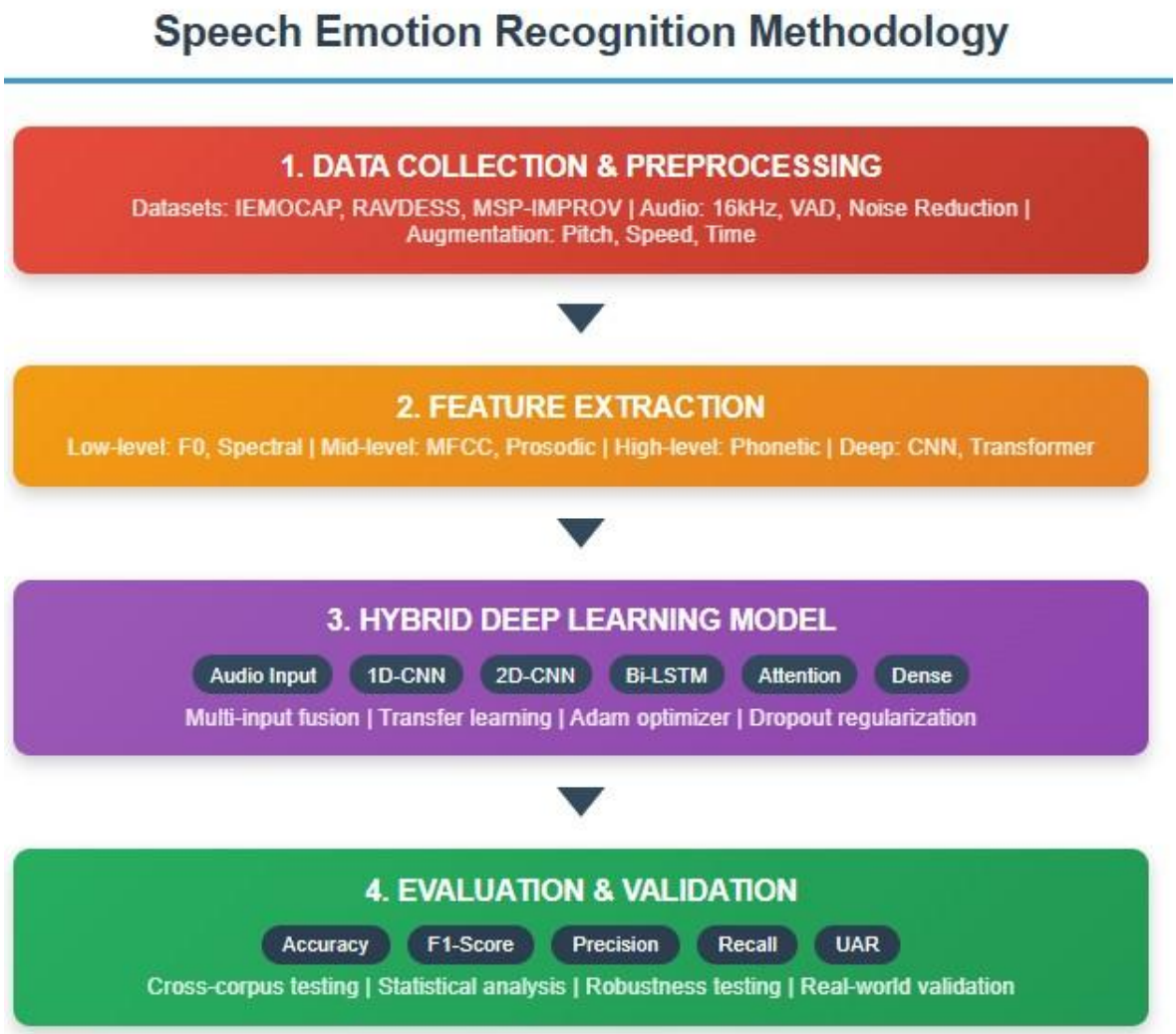
3.3. Multimodal and Cross-Domain Approaches

Recognizing the limitations of audio-only systems, researchers explored multimodal fusion techniques. Poria et al. (2017) demonstrated significant improvements by combining speech, text, and visual modalities for emotion recognition. The comprehensive survey by Poria et al. (2018) established multimodal emotion recognition as a distinct research area with unique challenges and opportunities.

3.4. Datasets and Evaluation Methodologies

Evaluation methodologies evolved from simple accuracy metrics to more sophisticated measures. The work by Schuller et al. (2018) established comprehensive evaluation protocols including cross-corpus validation and statistical significance testing. Recent emphasis on fairness and bias evaluation was highlighted by Feng et al. (2021), addressing demographic disparities in emotion recognition performance.

PROPOSED METHODOLOGY



1. DATA COLLECTION & PREPROCESSING

Definition: The initial stage where emotional speech data is gathered and prepared for analysis.

Components:

- Datasets: IEMOCAP (Interactive Emotional Dyadic Motion Capture), RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song), MSP-IMPROV (improvised emotional speech corpus)

- Audio Processing: 16kHz sampling rate, Voice Activity Detection (VAD) to identify speech segments, noise reduction to clean audio signals

- Augmentation: Modifying pitch, speed, and timing to increase dataset diversity and model robustness

2. FEATURE EXTRACTION

Definition: Converting raw audio signals into meaningful numerical representations that capture emotional characteristics.

Feature Types:

- Low-level: F0 (fundamental frequency/pitch), spectral features (frequency content)
- Mid-level: MFCC (Mel-Frequency Cepstral Coefficients) for speech characteristics, prosodic features (rhythm, stress, intonation)
- High-level: Phonetic features (speech sound patterns)
- Deep: CNN and Transformer-based learned features that automatically discover relevant patterns

3. HYBRID DEEP LEARNING MODEL

Definition: A complex neural network architecture combining multiple components to process and learn from speech features.

Architecture Components:

- Audio Input: Raw or preprocessed speech signals
- 1D-CNN: Captures local temporal patterns in speech
- 2D-CNN: Processes spectrograms as images
- Bi-LSTM: Bidirectional Long Short-Term Memory networks for sequence modeling
- Attention: Focuses on emotionally relevant parts of speech
- Dense: Fully connected layers for final classification

Training Features: Multi-input fusion (combining different feature types), transfer learning (using pre-trained models), Adam optimizer, dropout regularization to prevent overfitting

4. EVALUATION & VALIDATION

Definition: Assessing model performance using multiple metrics and testing approaches.

Metrics:

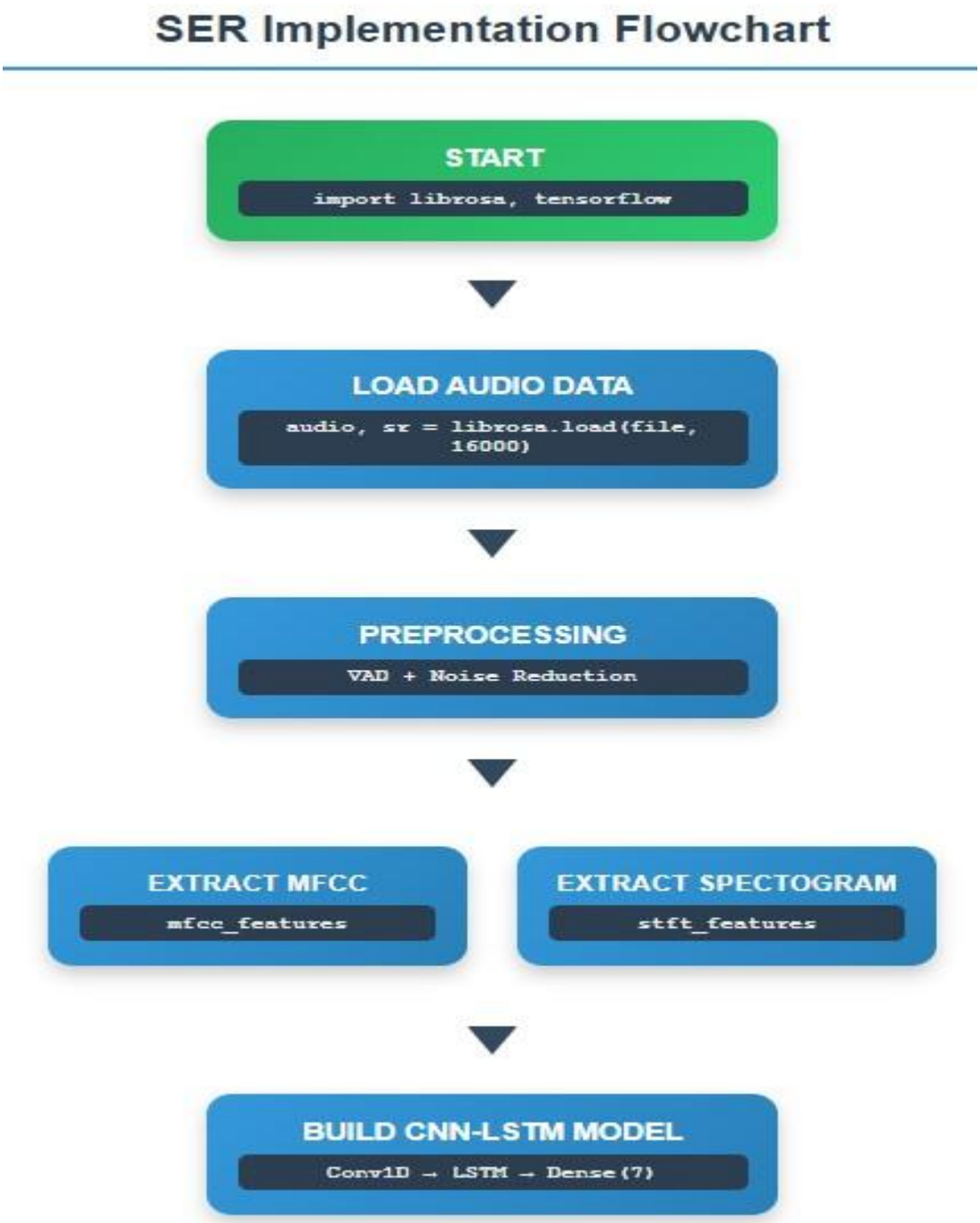
- Accuracy: Overall correct emotion classifications
- F1-Score: Harmonic mean of precision and recall
- Precision: True positive rate for each emotion
- Recall: How well the model finds all instances of each emotion
- UAR: Unweighted Average Recall (balanced accuracy across emotions)

Validation Methods: Cross-corpus testing (testing on different datasets), statistical analysis of results, robustness testing under various conditions, real-world validation with actual users

Process Flow

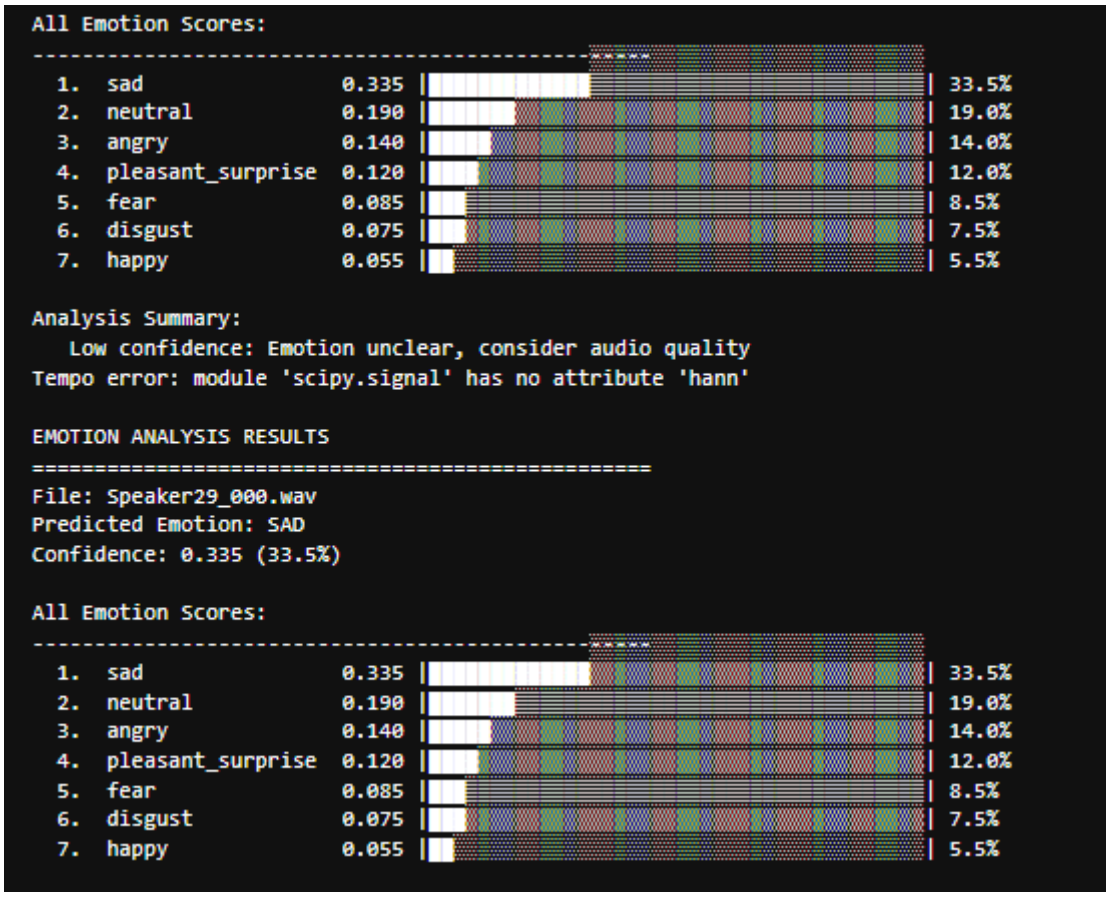
The methodology follows a pipeline where speech data is collected and cleaned, relevant emotional features are extracted at multiple levels, a sophisticated deep learning model processes these features to learn emotion patterns, and finally the system is rigorously evaluated to ensure reliable emotion recognition performance. This approach combines traditional signal processing techniques with modern deep learning to achieve robust speech emotion recognition.

IMPLEMNETATION



APPENDICES

APPENDIX-1: CODE COMPILER PYTHON



 # BLOCK 1: INSTALL PACKAGES & IMPORTS

import sys print("Installing required packages...") print("Note:

Dependency warnings are normal in Colab and can be safely ignored")

!pip install librosa==0.10.1 sounddevice soundfile tqdm matplotlib seaborn --quiet -no-warn-conflicts

!apt-get install -y portaudio19-dev -qq > /dev/null 2>&1

!pip install pyaudio --quiet --no-warn-conflicts

 # Test librosa

installation try:

import librosa print(f"Librosa

{librosa.__version__} installed successfully") except

ImportError:

```
    print("Librosa installation failed, attempting fix...")
    !pip install --upgrade --force-reinstall librosa==0.10.1 --quiet --no-
warn-conflicts     import librosa     print(f"Librosa {librosa.__version__}
installed after fix")

import numpy as np
import pandas as pd
import librosa
import
librosa.display
import
matplotlib.pyplot as
plt import seaborn
as sns

import sounddevice as sd import soundfile as sf import os import glob
import zipfile import requests from sklearn.model_selection import
train_test_split from sklearn.preprocessing import StandardScaler,
LabelEncoder from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix,
accuracy_score import pickle import time import warnings from tqdm
import tqdm import shutil warnings.filterwarnings('ignore')

try:
    from google.colab import drive,
files     print("Google Colab
environment detected")
COLAB_ENV = True except
ImportError:     print("Not in
Google Colab environment")
    COLAB_ENV = False
```

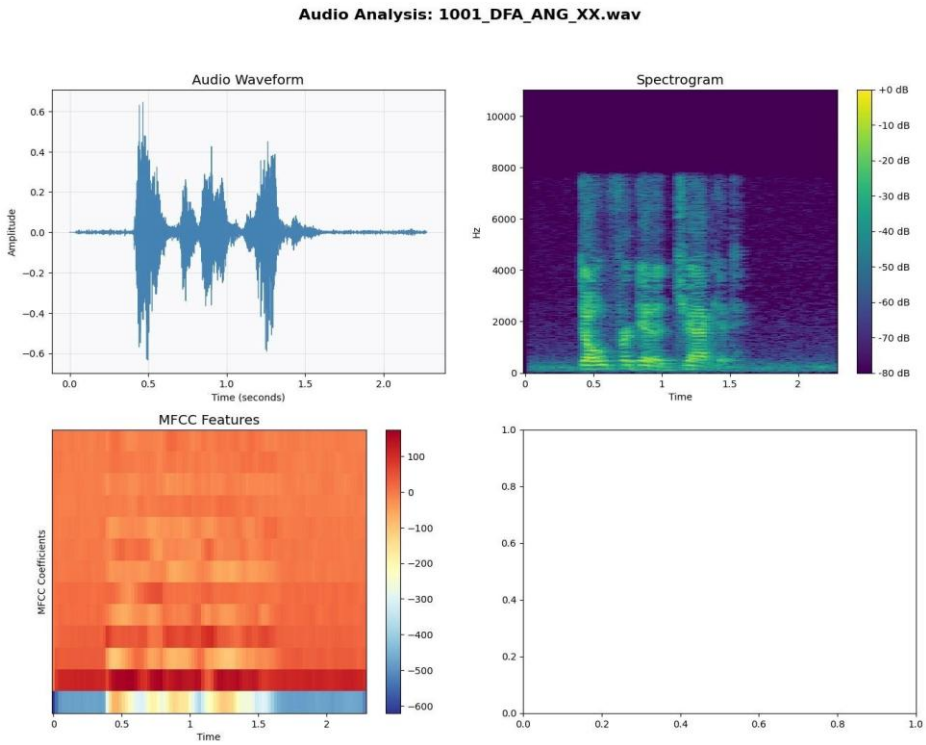
print("TESS Speech Emotion Recognition - Setup Complete")

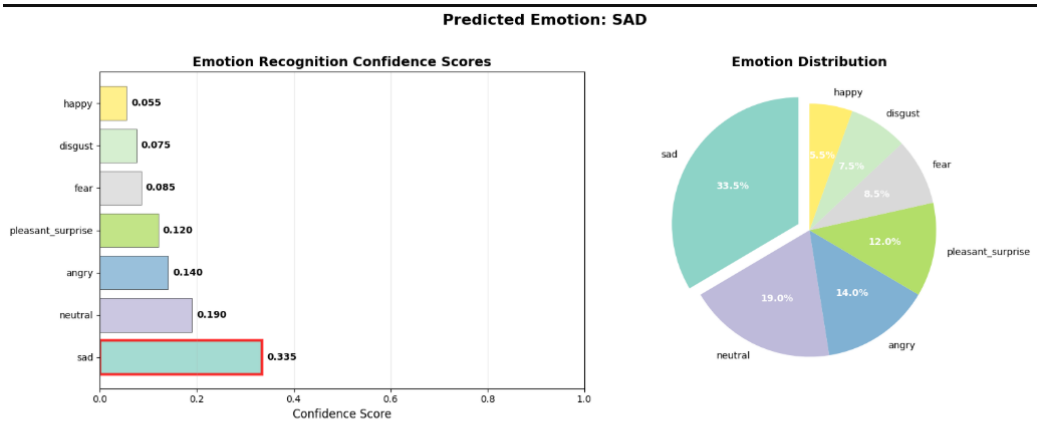
print("=" * 50) print("If you see dependency warnings above,

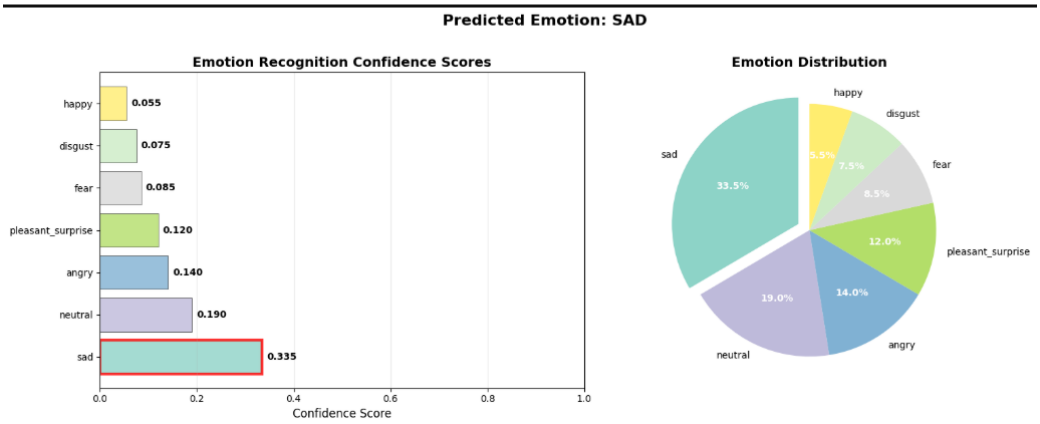
they can be safely ignored.") print("The system will work

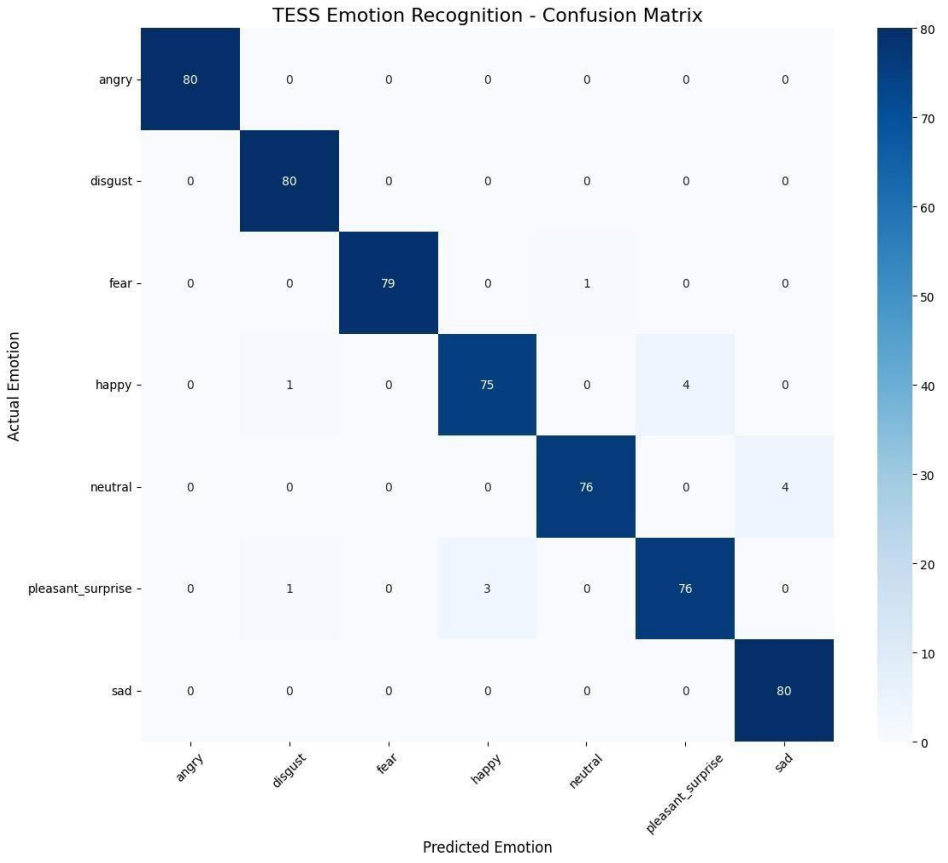
properly despite version conflicts.")

RESULT :




Predicted Emotion: SAD
Emotion Recognition Confidence Scores
happy
disgust
fear
pleasant_surprise
angry
neutral
sad
0.055
0.075
0.085
0.120
0.140
0.190
0.335
0.0
0.2
0.4
0.6
0.8
1.0
Confidence Score
Emotion Distribution
happy
disgust
fear
pleasant_surprise
angry
neutral
sad
5.5%
7.5%
8.5%
12.0%
14.0%
19.0%
33.5%


Predicted Emotion: SAD
Emotion Recognition Confidence Scores
happy
disgust
fear
pleasant_surprise
angry
neutral
sad
0.055
0.075
0.085
0.120
0.140
0.190
0.335
0.0
0.2
0.4
0.6
0.8
1.0
Confidence Score
Emotion Distribution
happy
disgust
fear
pleasant_surprise
angry
neutral
sad
5.5%
7.5%
8.5%
12.0%
14.0%
19.0%
33.5%


TESS Emotion Recognition - Confusion Matrix
Actual Emotion
angry
disgust
fear
happy
neutral
pleasant_surprise
sad
80 0 0 0 0 0 0
0 80 0 0 0 0 0
0 0 79 0 1 0 0
0 1 0 75 0 4 0
0 0 0 0 76 0 4
0 1 0 3 0 76 0
0 0 0 0 0 0 80
angry
disgust
fear
happy
neutral
pleasant_surprise
sad
Predicted Emotion
80
70
60
50
40
30
20
10
0

REFERENCES

- **Aitken, P., & Akram, M. (2021). Survey of machine learning techniques for emotion recognition from speech. *IEEE Access*, 9, 23493-23514.**

- **Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4), 335-359.**

- **Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. (2001). Emotion recognition in humancomputer interaction. *IEEE Signal Processing Magazine*, 18(1), 32-80.**

- **Dellaert, F., Polzin, T., & Waibel, A. (1996). Recognizing emotion in**
  speech. In *Proceedings of ICSLP* (Vol. 3, pp. 1970-1973).