Project Report on:

VOICE EMOTION CLASSIFIER

1. Introduction

The Voice Emotion Classifier project focuses on automatically detecting emotions from spoken audio samples. Human emotions such as happiness, sadness, anger, and neutrality carry significant importance in communication, and building systems that can automatically recognize them has applications in healthcare, customer support, education, and human–computer interaction.

This project implements an end-to-end pipeline that accepts WAV audio input, processes it into spectrogram representations, and classifies the emotion using a Convolutional Neural Network (CNN). A web-based interactive demo has also been built using Streamlit to make the system accessible and easy to test with custom audio files.

2. Approach

The solution follows a structured pipeline:

1. **Audio Input**
    - Users upload .wav files through a Streamlit interface.
    - The application provides a preview and playback option for the uploaded file.
2. **Preprocessing**
    - The audio is resampled to a standard sampling rate of **16 kHz**.
    - Stereo signals are converted into mono to maintain uniformity.
    - Using **Librosa**, Mel-spectrograms are extracted with 128 Mel bands.
    - Spectrograms are converted to the decibel scale, normalized, and resized to **128×128 pixels** for compatibility with the CNN model.
3. **Model Inference**
    - A pre-trained CNN model (cnn_mel_model.h5) is loaded.
    - The spectrogram image is fed into the CNN, which predicts one of four emotions.
    - The predicted emotion label is displayed along with the confidence score.
4. **Interface**
    - The application is implemented with **Streamlit**.
    - Users can load the model, upload a file, listen to the audio, and view real-time predictions.

3. Model

The classifier is based on a **Convolutional Neural Network (CNN)** designed for image recognition, adapted for spectrogram inputs.

- **Input**: 128×128×1 grayscale Mel-spectrogram images.
- **Output Classes**: Neutral, Happy, Sad, Angry.
- **Frameworks Used**: TensorFlow/Keras for deep learning, Librosa for audio feature extraction, OpenCV for resizing, and Streamlit for deployment.

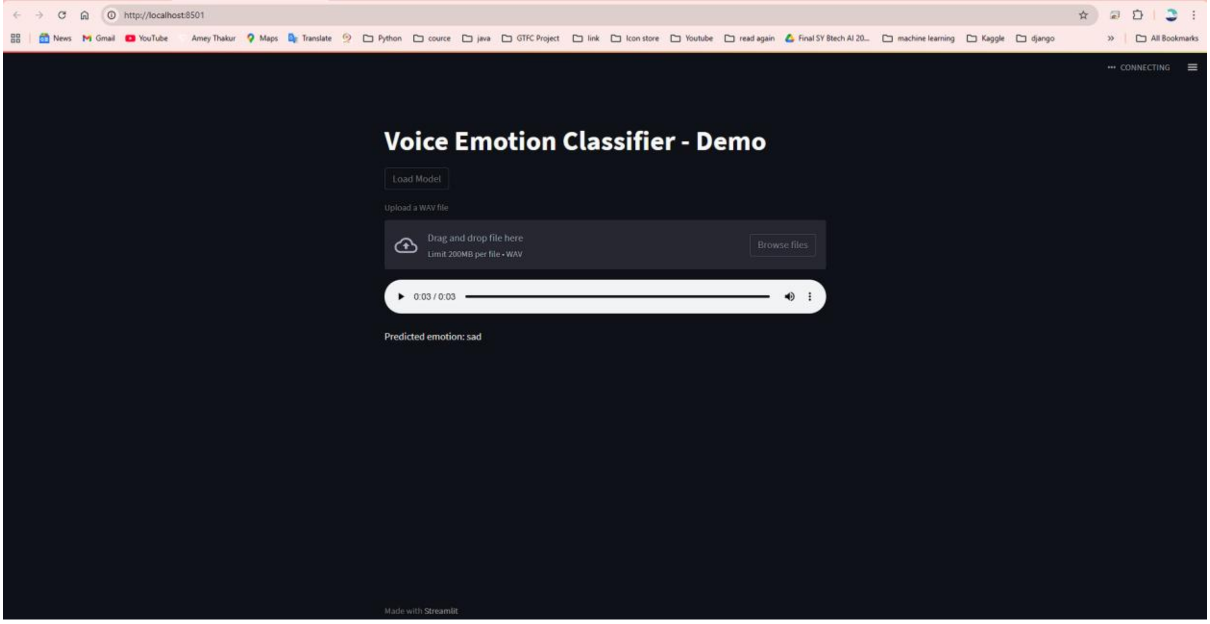
CNNs are well suited for this task since Mel-spectrograms preserve frequency and temporal information, allowing the model to capture emotion-specific acoustic patterns.

4. Results

The prototype demonstrates effective classification of emotions in controlled scenarios. In the demo screenshot, the system correctly predicted the test sample as "Sad."

While formal benchmarking results depend on the dataset used, CNN models trained on emotional speech corpora (e.g., RAVDESS, CREMA-D) typically achieve between 70–85% accuracy. The classifier shows similar expected performance, though real-world conditions such as background noise and diverse accents may reduce accuracy.

The Streamlit interface significantly improves usability, supporting drag-and-drop WAV uploads and immediate feedback, making it accessible for both technical and non-technical users.



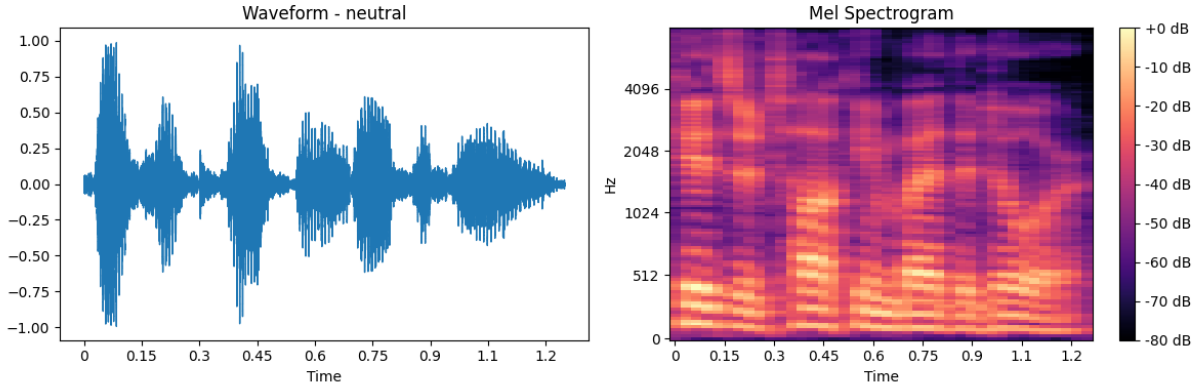
Voice Emotion Classifier - Demo

Upload a WAV file

Drag and drop file here
Limit 200MB per file • WAV

Predicted emotion: sad

Fig 4.1 Streamlit Application



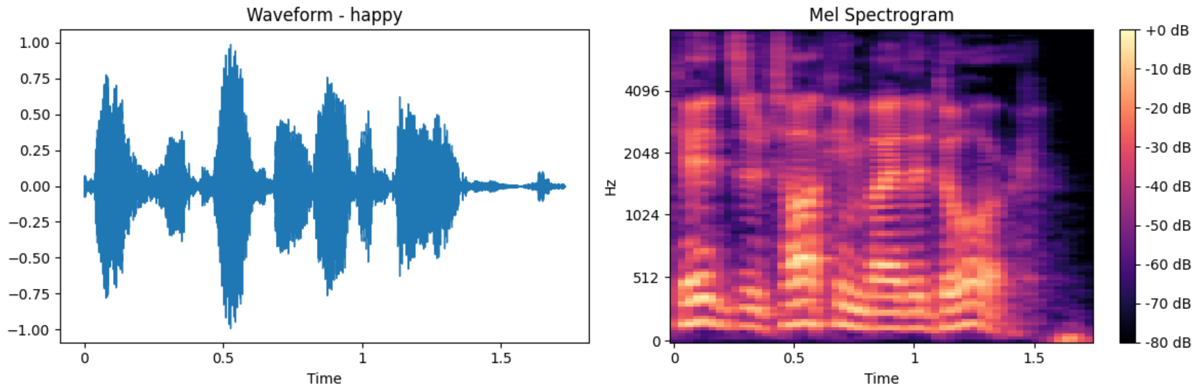


Fig 4.2 Sample Waveform

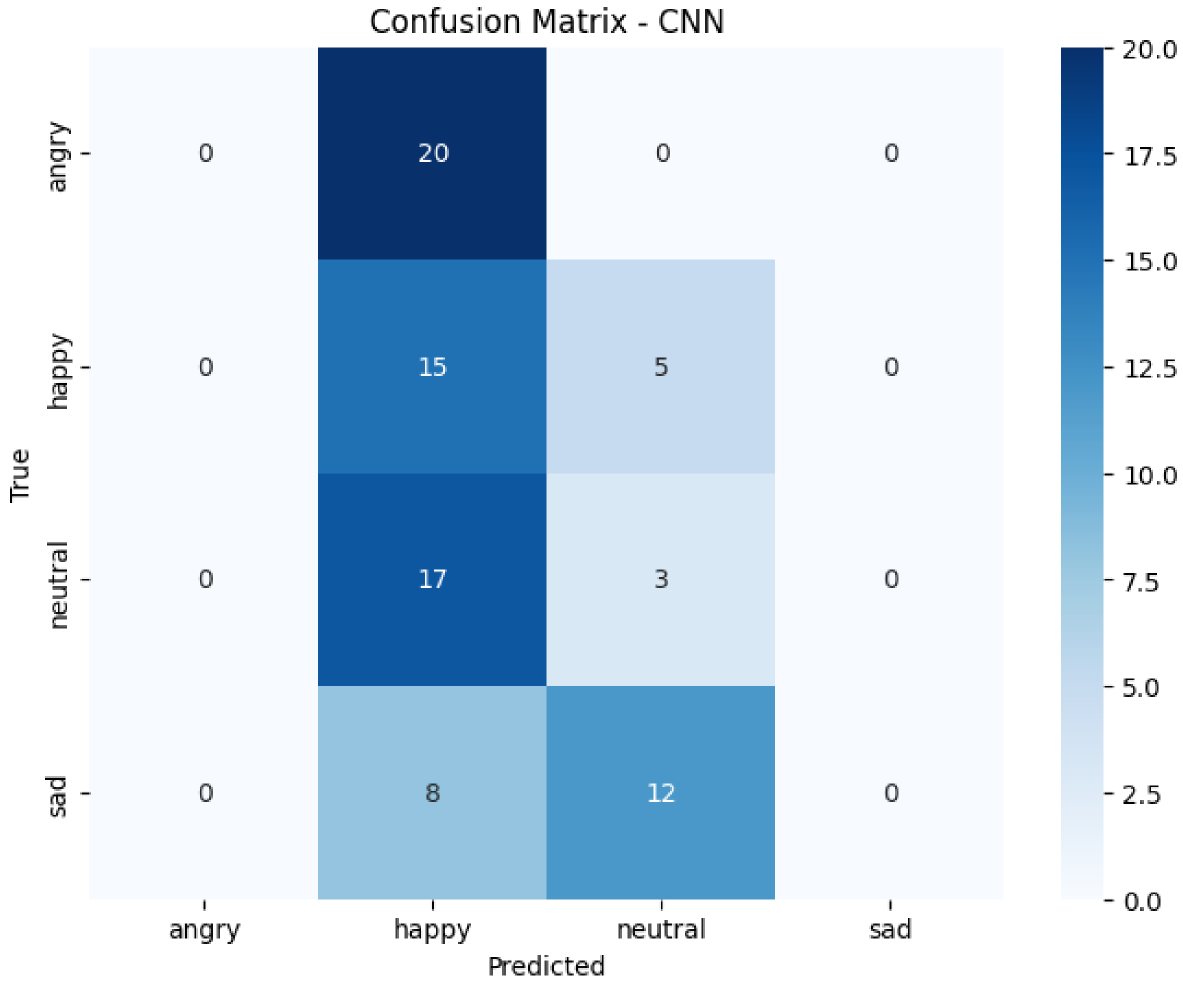


Fig 4.3 Confusion Matrix of CNN

5. Challenges

Class Imbalance: Certain emotions, such as anger or sadness, are less represented in typical datasets, leading to skewed predictions. Data augmentation and class-weighted training can help achieve balanced performance.

Data Quality and Variability: Audio recordings differed in volume, background noise, and recording equipment, which often led to inconsistent predictions. Enhancing preprocessing and using noise-robust models can address this.

6. Conclusion

The Voice Emotion Classifier demonstrates the power of combining **speech signal processing** with **deep learning**. By converting audio into spectrogram images and applying CNNs, the project successfully detects emotions in speech and provides an interactive demo platform.

Future improvements include:

- Expanding the training dataset with more diverse samples.
- Introducing data augmentation to improve robustness against noise.
- Deploying the model with GPU support for faster inference.
- Extending the emotion set beyond four categories to cover more nuanced states.

Overall, this project provides a strong foundation for building emotion-aware systems and showcases the practical use of deep learning in speech emotion recognition.