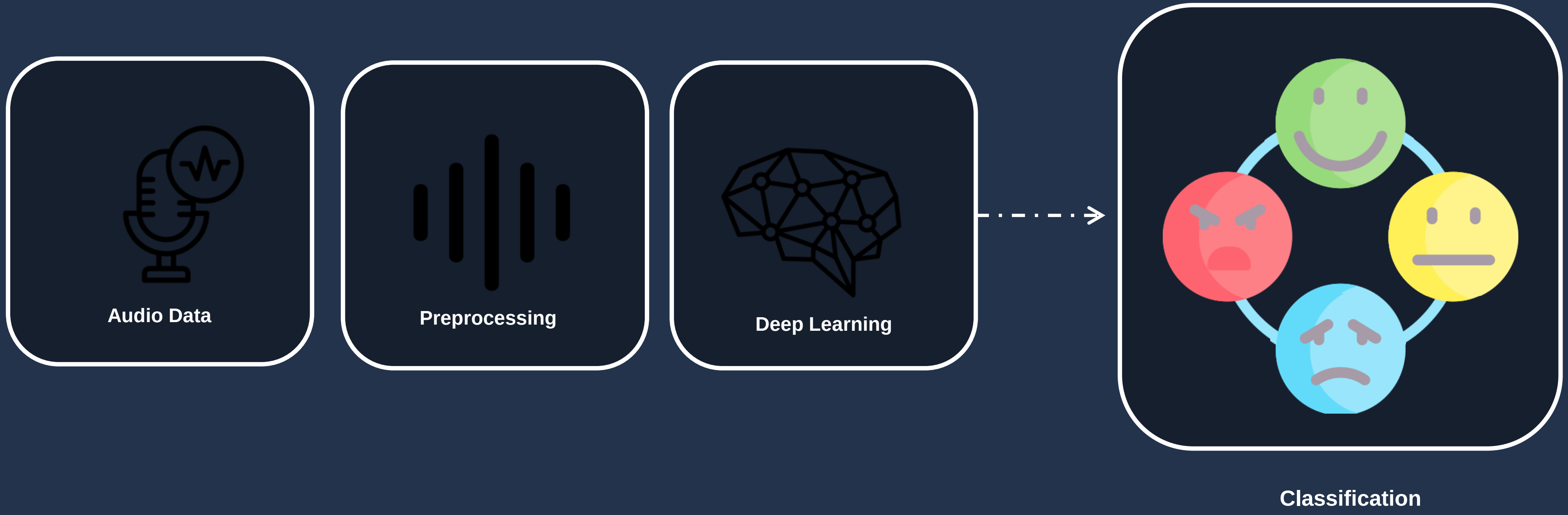
Voice Sentiment Analysis

Ahmed AbdElhamid Shenawy

Table of Contents

Problem Definition

Voice Sentiment Analysis has emerged as a vital area of research in affective computing and machine learning due to its diverse applications in human-computer interaction, mental health monitoring, and customer service systems. Human speech carries rich paralinguistic information that can be analyzed to infer emotional states. Key features such as pitch, intensity, and Mel spectrograms enable the quantitative representation of these emotional cues.



Problem Statement

- Why Sentiment Analysis Matters:
  - Helps businesses understand customer emotions.
  - Enhances human-computer interactions.
  - Valuable for mental health monitoring and social analytics
- Objective:
  - Develop a deep learning model to classify emotions (angry, happy, sad, neutral) from voice recordings.
- Challenges:
  - Limited dataset size.
  - Data preprocessing and augmentation challenges.

Solution Overview

- Approach:
    - Preprocess the audio data.
    - Augment the dataset to improve performance.
    - Feature extraction using mel-spectrograms.
    - Implement a CNN-based model for classification.
- Dataset :
    - The **EYASE** dataset used in this study contains audio recordings categorized by gender (male and female) in Arabic and their emotional states (happy, neutral, sad, and angry).

Dataset Characteristics

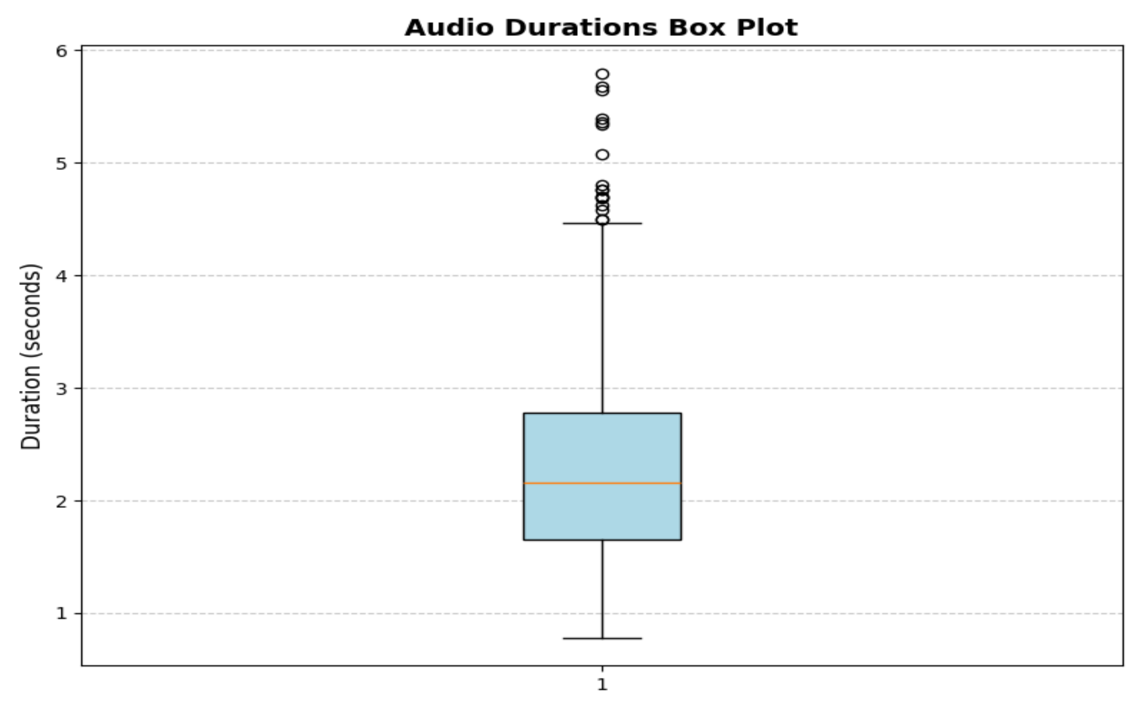
Total Files : 579
- Male Files : 339
- Female Files : 240

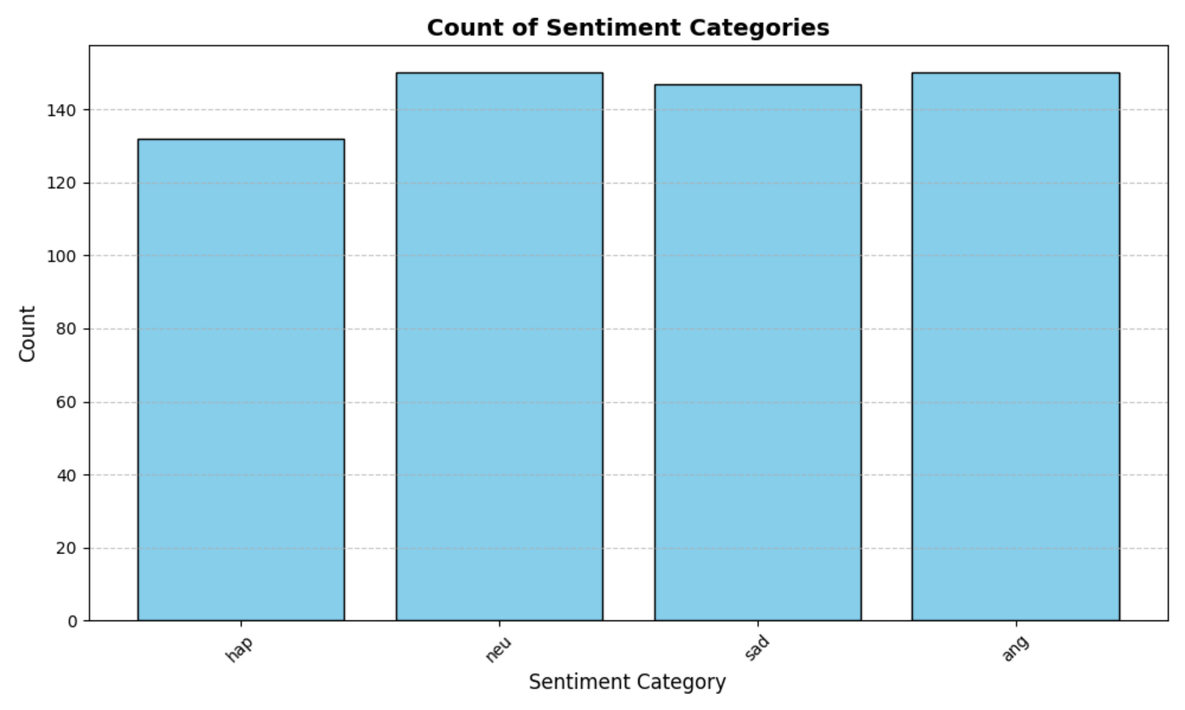
Sentiments:
- Sad: 147
- Happy: 132
- Angry: 150
- Neutral: 150

AVG Sample Rate: 44.1 KHz
AVG Audio Duration: 2.33 s

Data Exploration

- Leveraged data exploration techniques to analyze and visualize the distribution of the EYASE dataset, emphasizing the critical role of Exploratory Data Analysis (EDA) in uncovering key insights and shaping data-driven decisions





EDA graphs samples

Methods

Data Preprocessing:

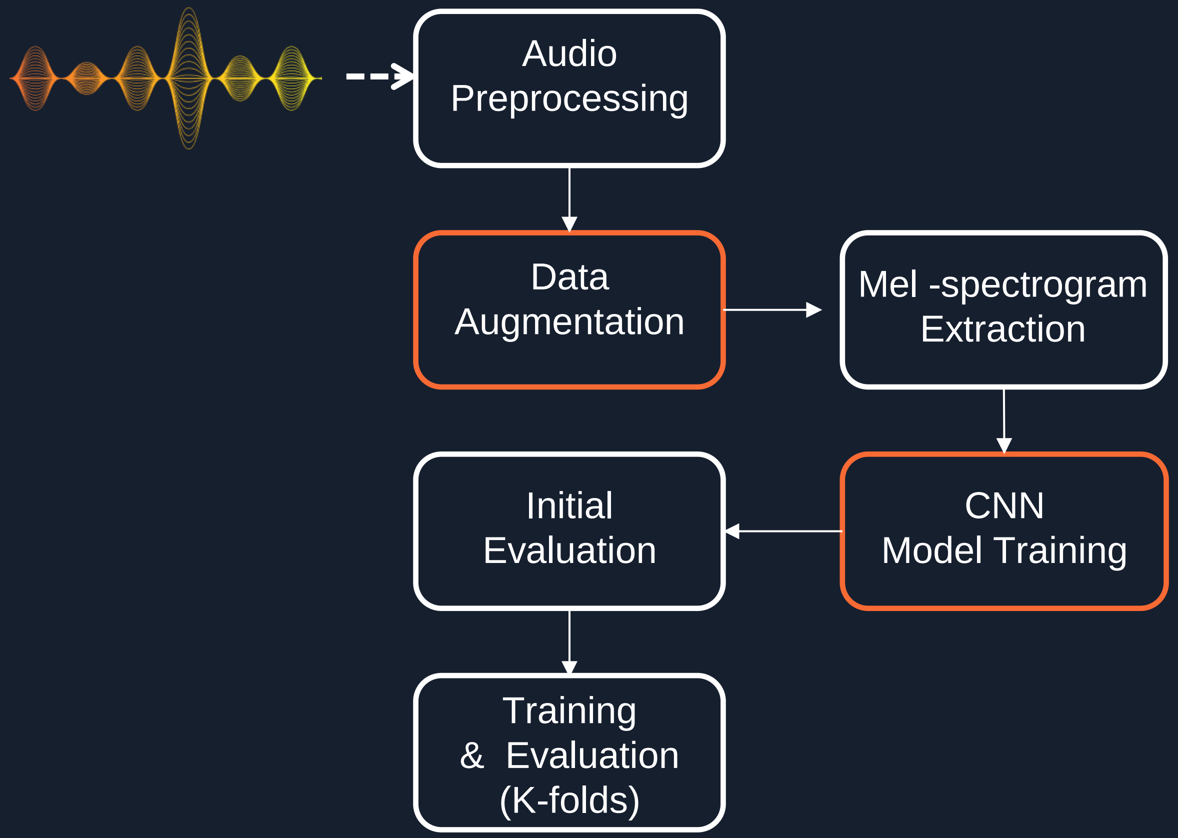
- Padding and truncating audio files to match the mean duration of the dataset.

Data Augmentation:

- Changing the pitch of the audio files.
- Adding white noise to create synthetic data.

Feature Extraction:

- Conversion of audio signals to **Mel Spectrograms**.
- Down sampling from 44100 Hz to 8000 Hz.
- Normalization of spectrogram values.



Methods Cont.

Data Truncating and Padding in detail:

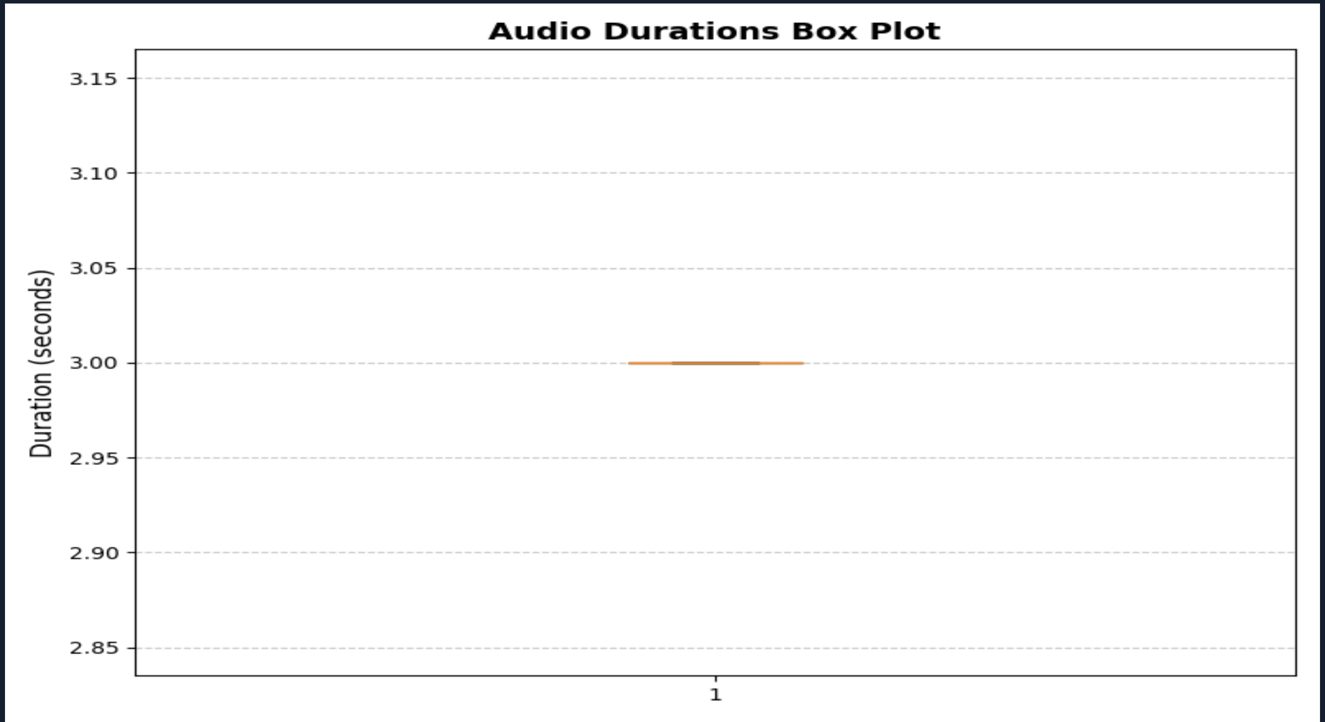
1. Padding

- Padding involves extending shorter audio signals to a fixed length by adding zeros (silence) at the end .

2. Truncating

- Truncating involves cutting off longer audio signals to fit a predefined length.

- Calculations done WRT Mean of audio samples



Methods Cont.

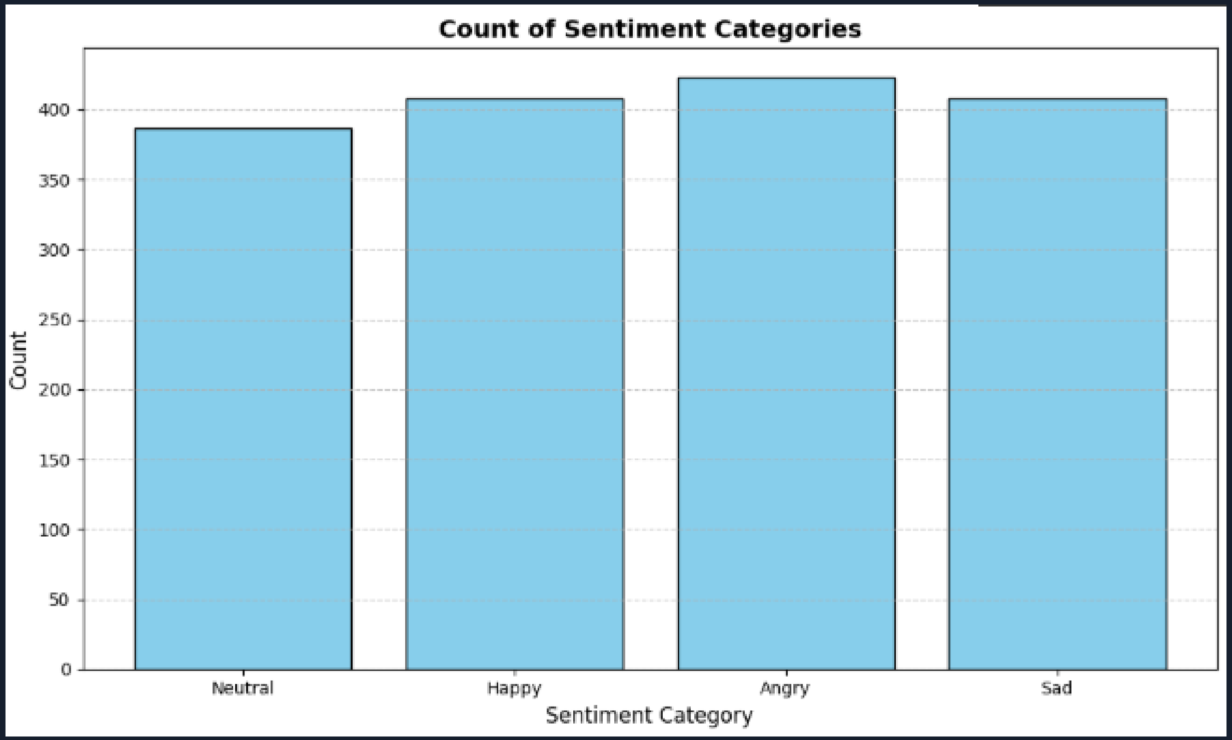
Data Augmentation in details:

1. Function: add_noise

This function adds random noise to an audio signal. The noise simulates a slight background disturbance, making the model robust to real-world audio variations.

2. Function: pitch_shift

This function shifts the pitch of the audio signal. Shifting the pitch simulates different voice characteristics (e.g., masculinizing or feminizing audio).

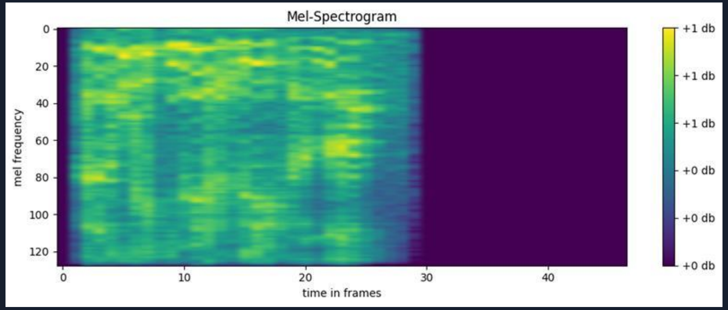


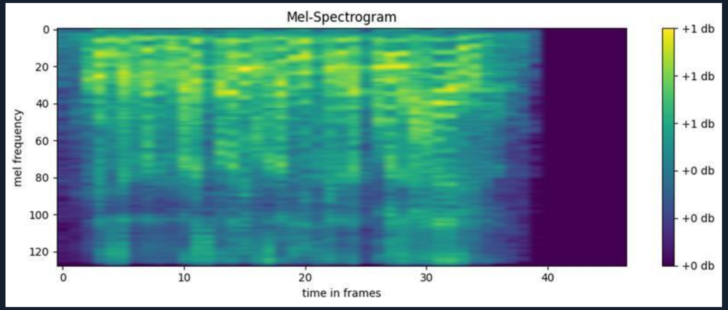
Total Files after Augmentation: **1626**

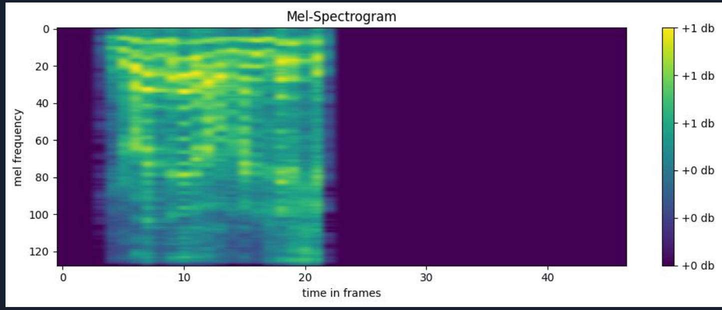
Mel-Spectogram

Mel-Spectogram: Representation in time frequency domain at Mel-Scale

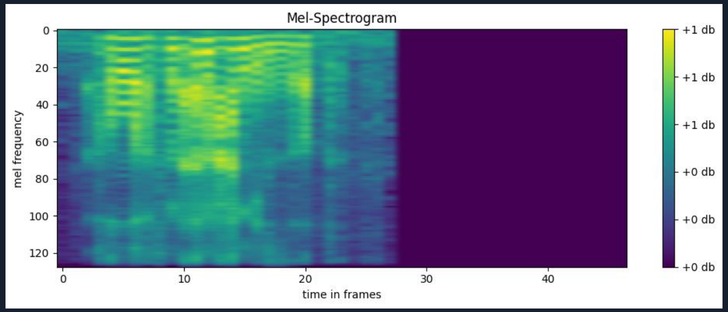
Angry Mel-Spectrogram

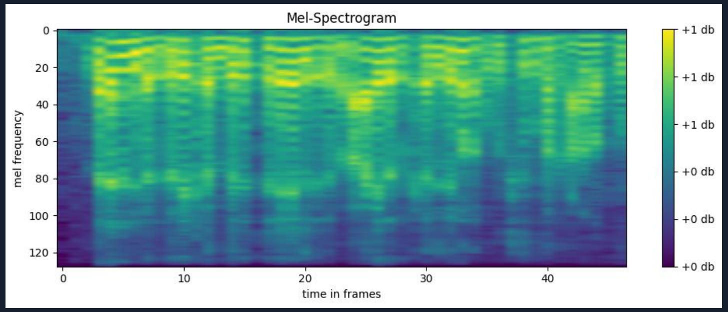


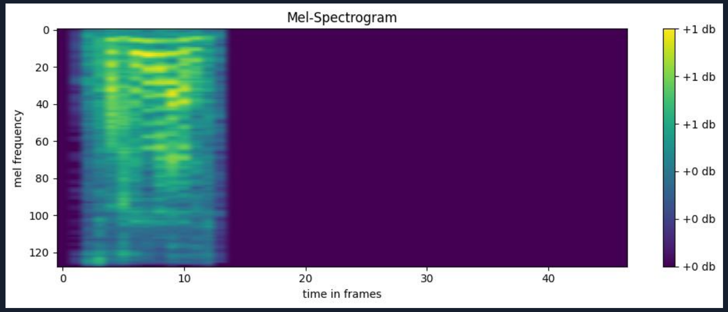




Neutral Mel-Spectrogram

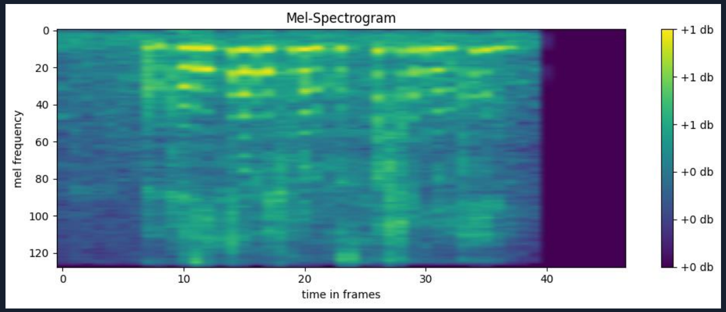


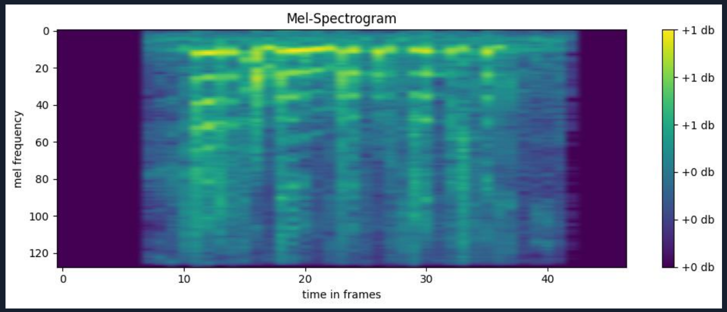


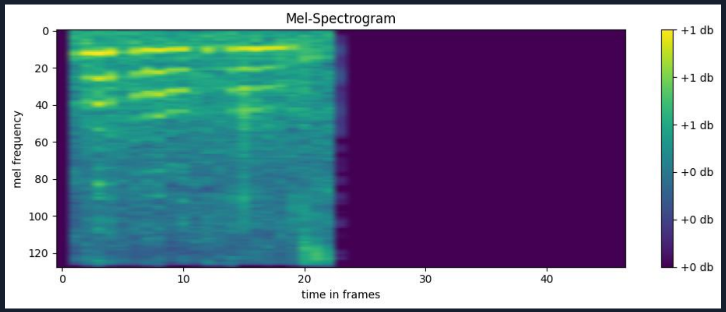


Mel-Spectogram

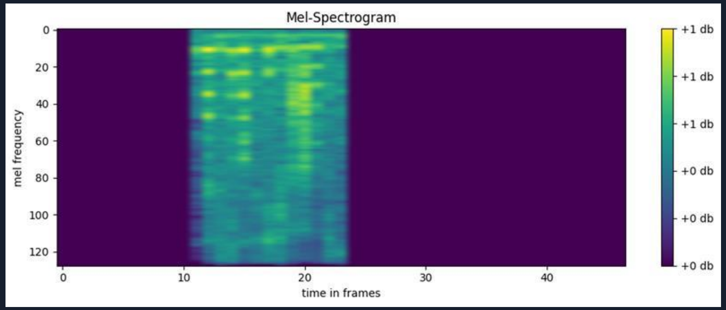
Sad Mel-Spectrogram

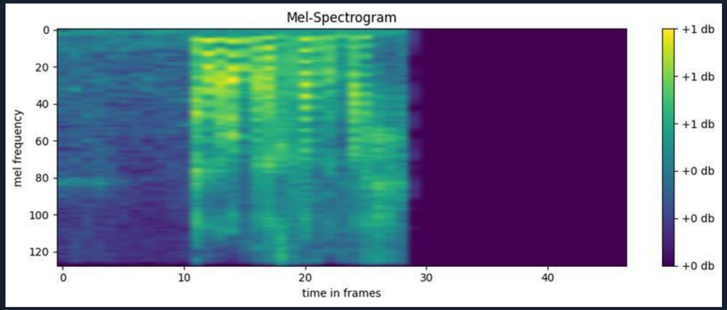


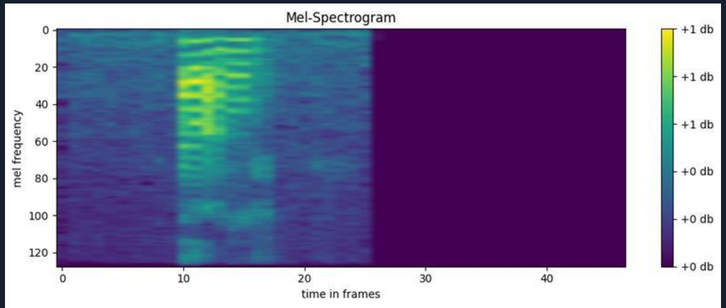




Happy Mel-Spectrogram







Methods Cont.

Data Splitting:

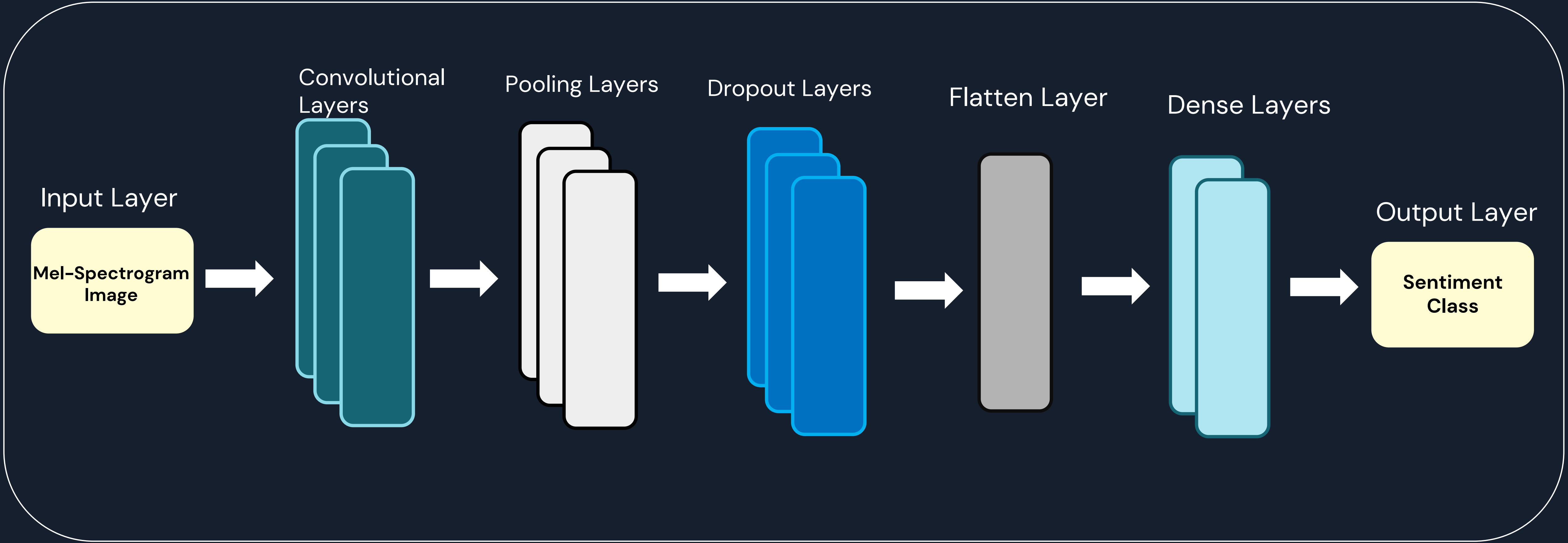
- Training set: 80%
- Testing set: 20%

Model Architecture:

- Convolutional Neural Network (CNN) with 12 layers:
    - Convolutional Layers (Conv2D): 3 layers → spatial features.
    - Pooling Layers (MaxPooling2D): 3 layers → Reduce dimensions.
    - Dropout Layers: 3 layers → Prevent overfitting.
    - Flatten Layer: 1 layer → Convert 2D to 1D.
    - Dense Layers (Fully Connected): 2 layers → Final predictions.

Methods Cont.

CNN Model



Results

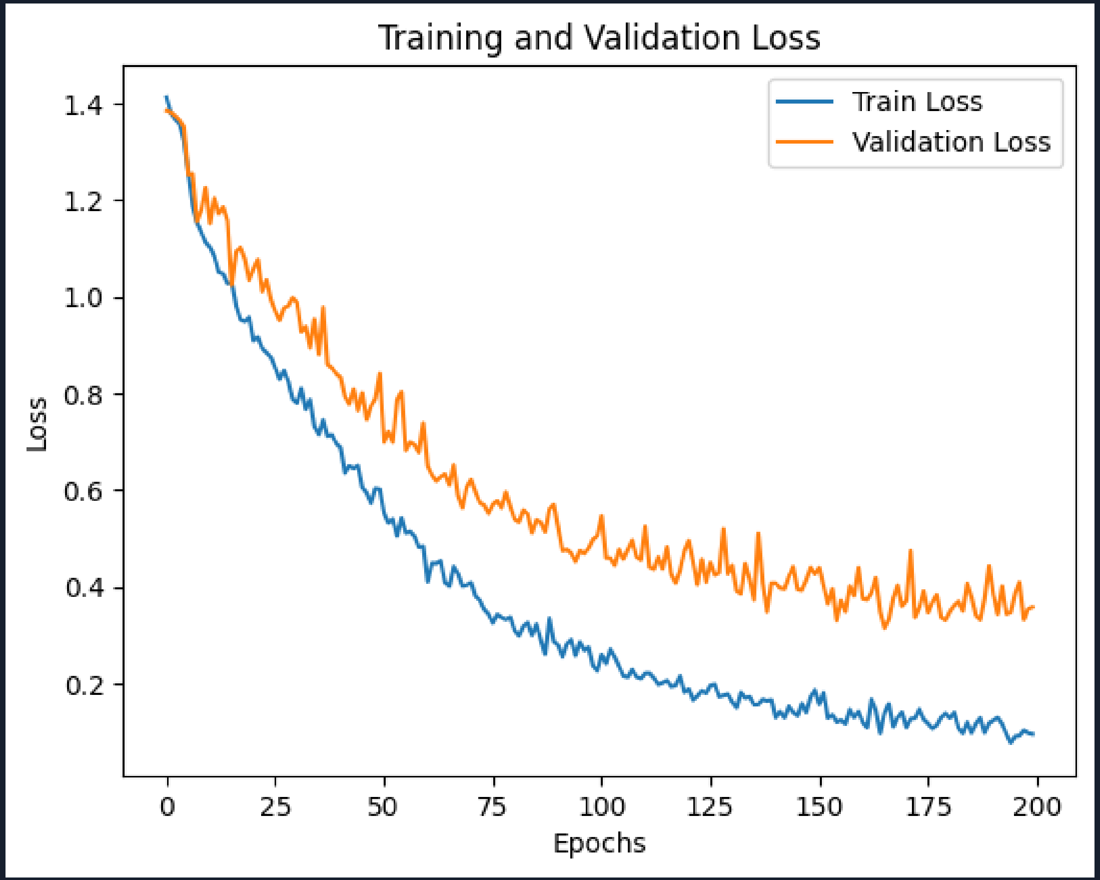
Initial Model Performance:

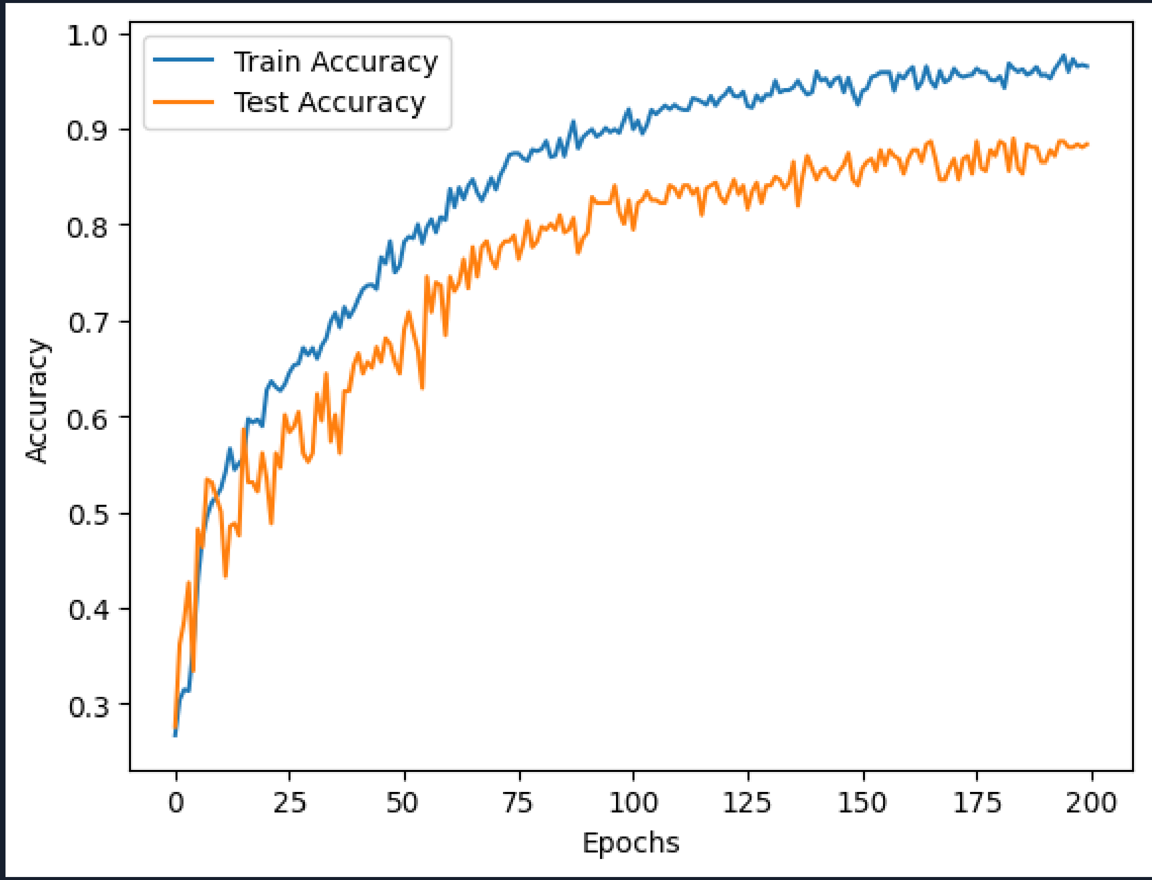
- **Accuracy:** 0.8676%
- **F1 Score:** 0.8678%
- **Recall : 0.8676%**
- **Precision : 0.8689%**

Actual / Predicted	Angry	Neutral	Happy	Sad
Angry	72	2	3	2
Neutral	2	72	4	11
Happy	1	4	63	0
Sad	0	10	4	75

Confusion matrix

Results Cont.





Validation

K-Fold Cross-Validation (5 Folds):

- **Avg Accuracy:** 0.8770%
- **Avg F1 Score:** 0.87626%
- **Avg Recall** : 0.8770%
- **Avg Precision :** 0.8792%

Conclusion

- Successful sentiment classification from voice data.
- Data augmentation improved model generalization.
- Future work includes exploring additional feature extraction techniques and optimizing model architecture.

Q&A

Thank you!