Speech Emotion Recognition using Traditional Machine Learning Techniques.

Umme Athiya

Introduction

What is Speech Motion Recognition?

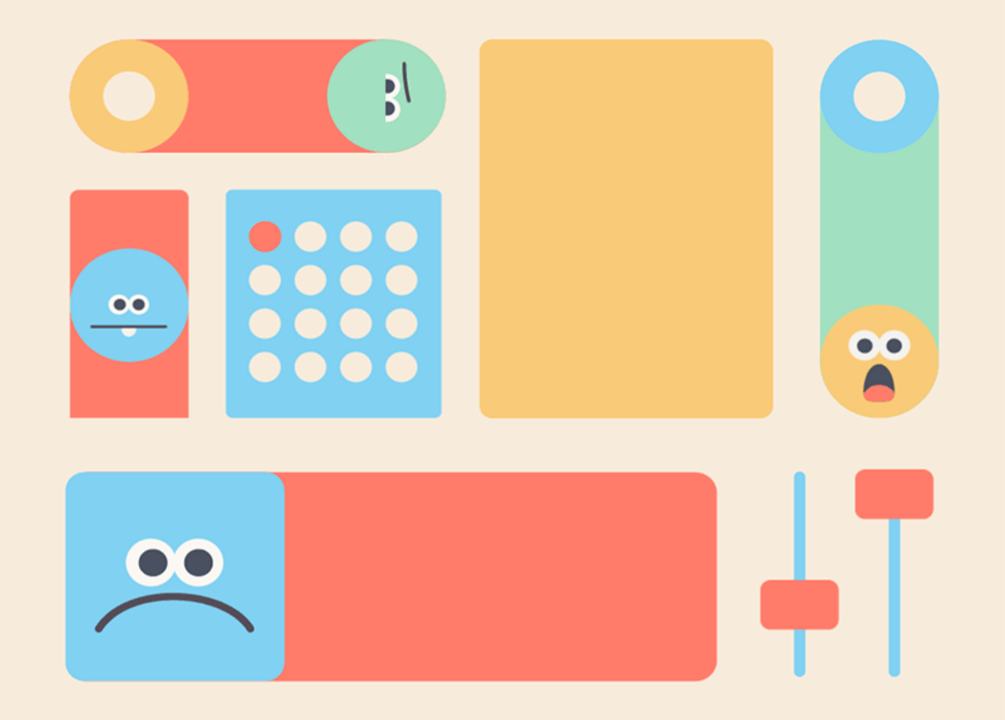
- An important task of actual identification and classification of motion patterns captured in the speech.
- It is the process of analyzing the waveforms and vibrations from the speech that aid in correctly classifying the patterns into suitable categories.

Why use Traditional Machine Learning?

- Generally, these ML models are efficient and highly interpretable.
- It can be implemented on structured datasets.

Why It Matters?

- **Healthcare**: Helping patients suffering from speech impairments.
- **Security:** It is useful in the domain of forensic investigations.
- **Human-Computer Interaction:** Elevating the effectiveness of voice-controlled devices.
- **Education:** Helping children and adults with speech therapy.



Dataset

- **<u>Sources of Data</u>**:

❖ Publicly available dataset from Kaggle.

- **<u>Features in the dataset:</u>**

❖ **<u>Spectral Features</u>**: these involves MFCCs, Spectrograms, Zero-Crossing Rate.

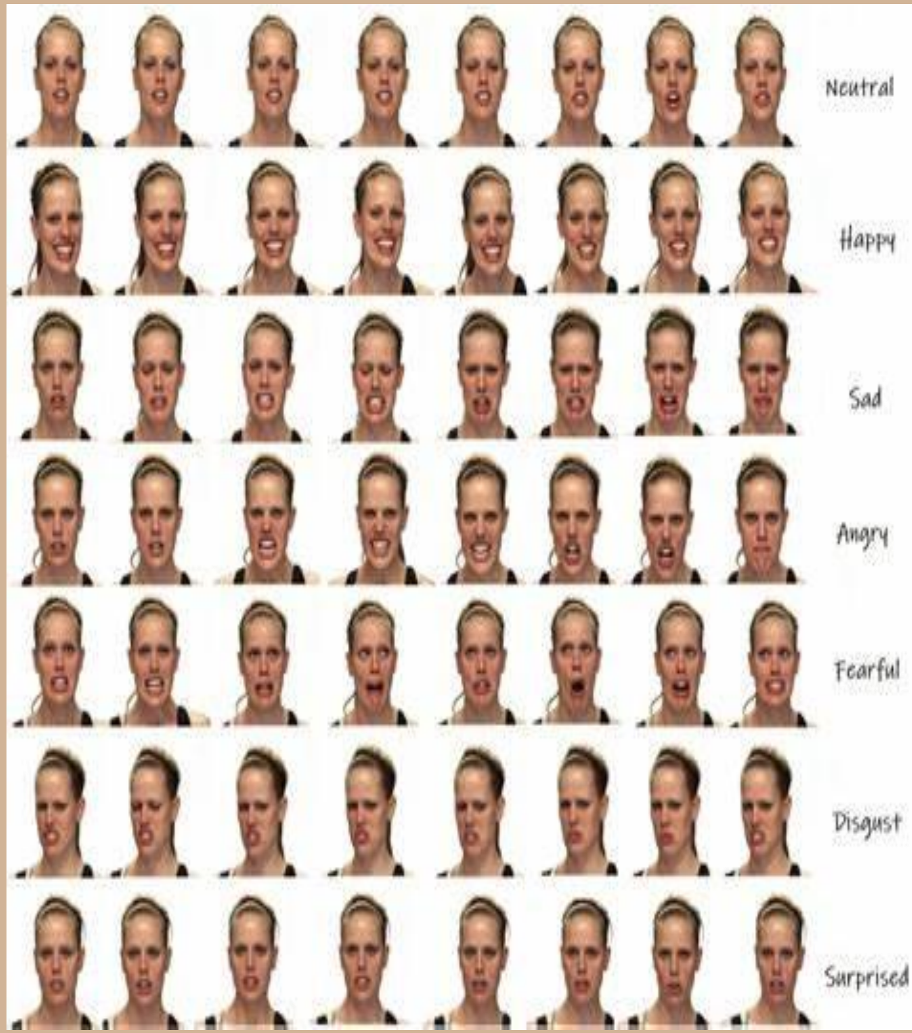
❖ **<u>Temporal Features:</u>** these involve energy, duration, pitch variation.

- **<u>Dataset Statistics:</u>**

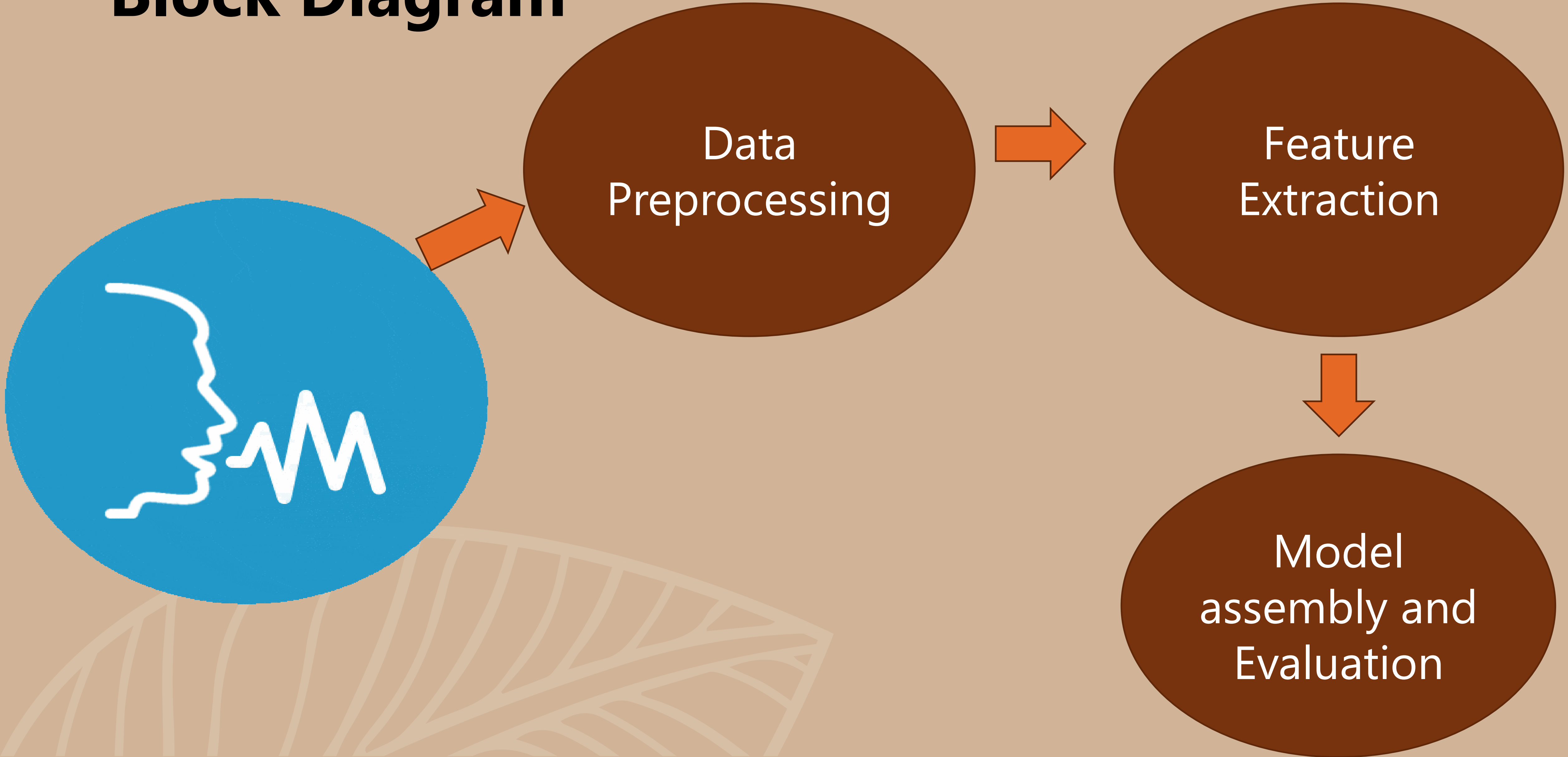
❖ 1440 audio files.

❖ 60 trials per actor – 24 professional actors (12 females & 12 males)

❖ 7 different emotional states – angry, happy, sad, neutral, fearful, surprised, and disgust.

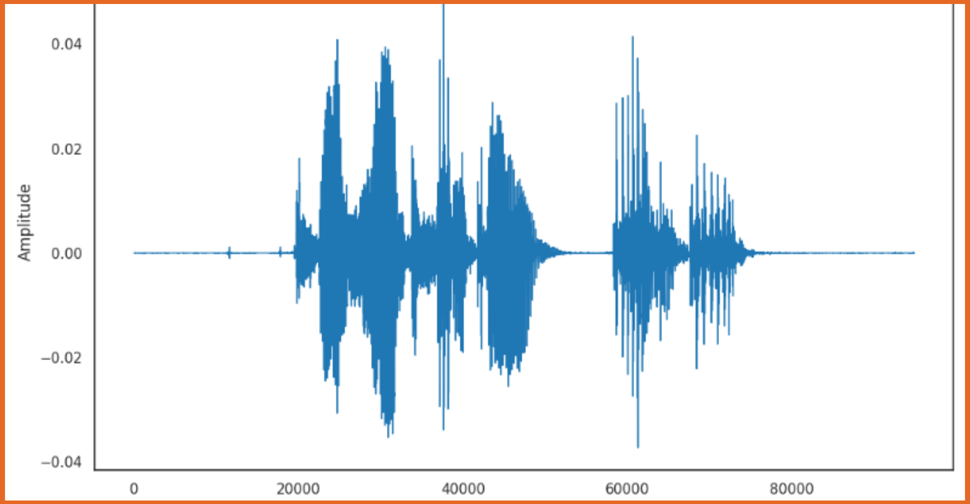


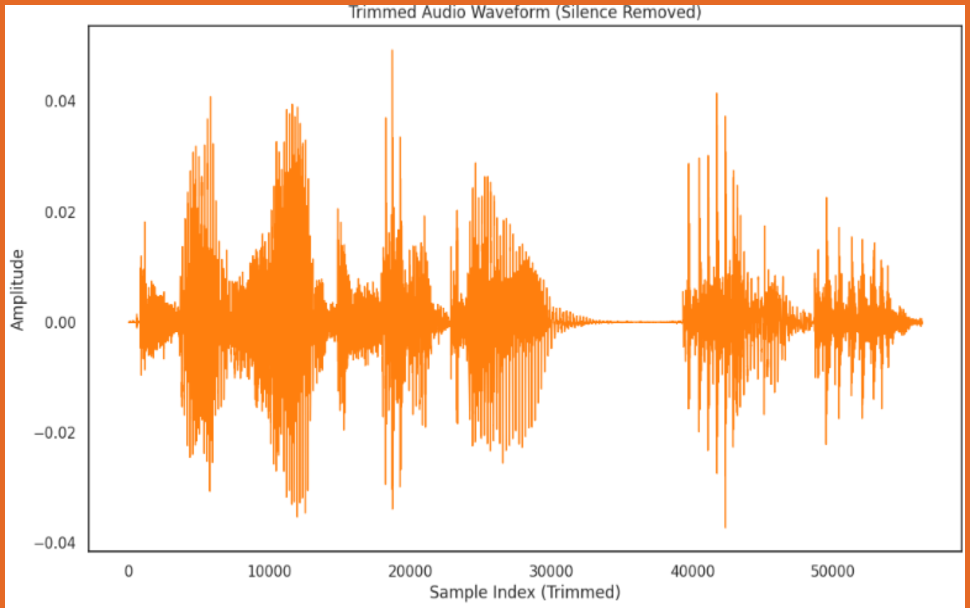
Block Diagram



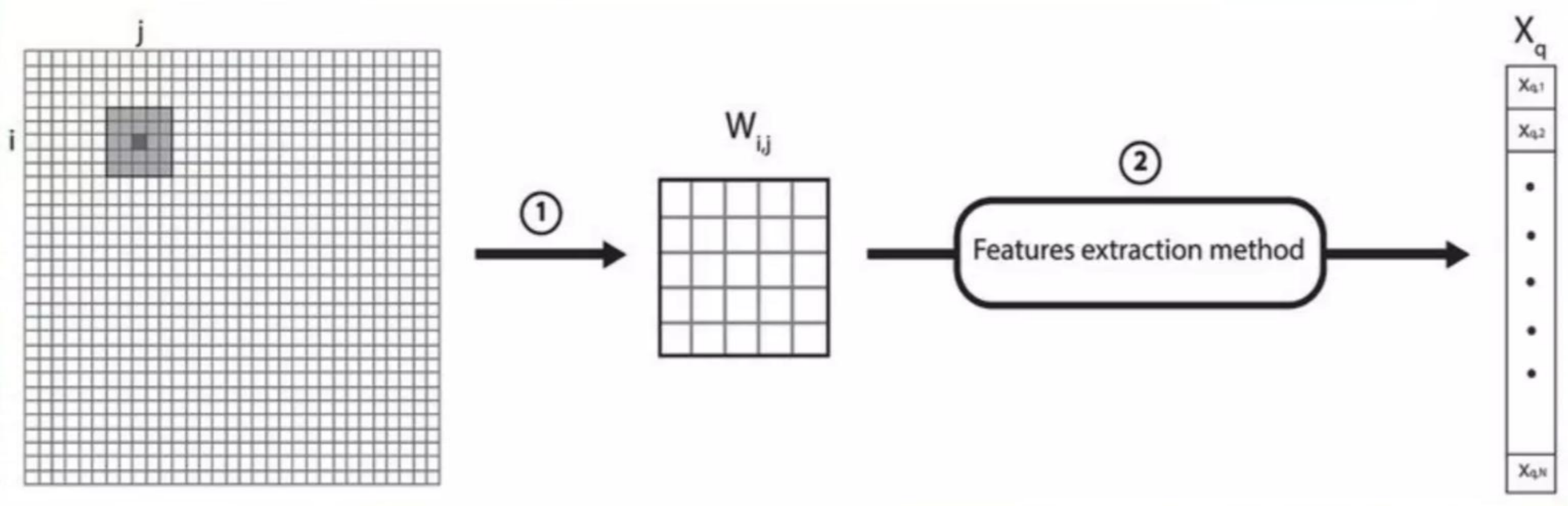
Data Preparation

- **Noise Reduction** – remove background noise.
- **Segmentation – splitting into lesser time frames.**
- **Normalization** – improve consistency.
- **Feature Engineering** – extracting features from frequency domain.





Feature Extraction

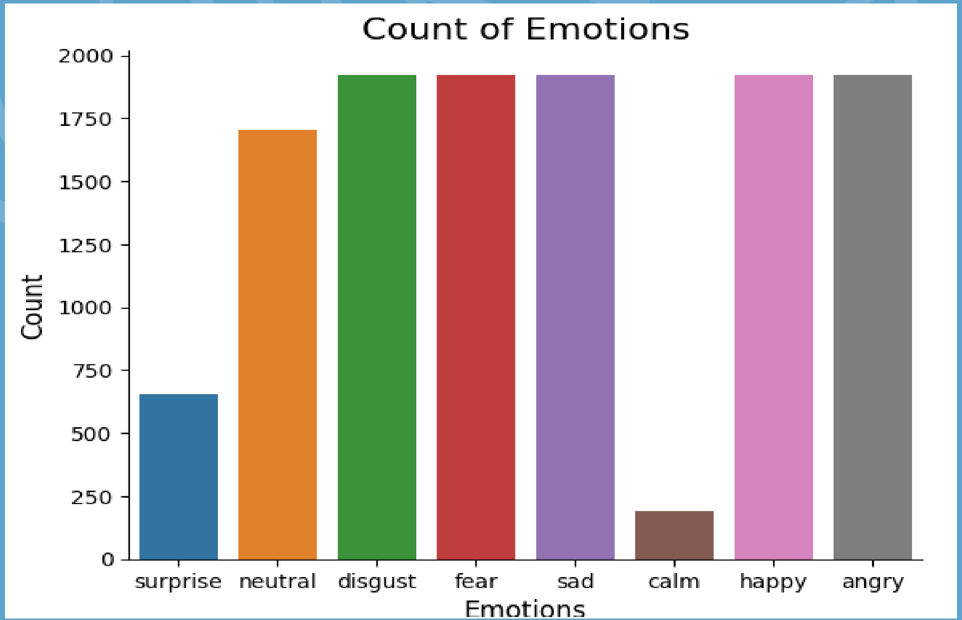


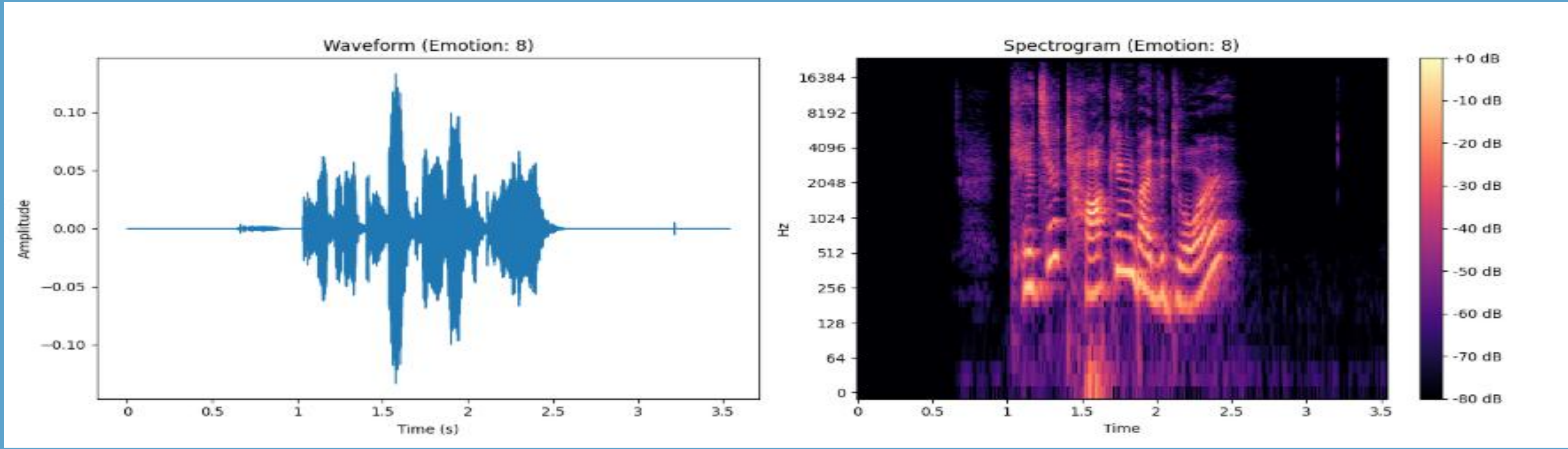
Mel- Frequency Cepstral Coefficients – it captures the perceptual features in the speech.

Spectrograms – generates a visual representation of any given frequency of speech over the time.

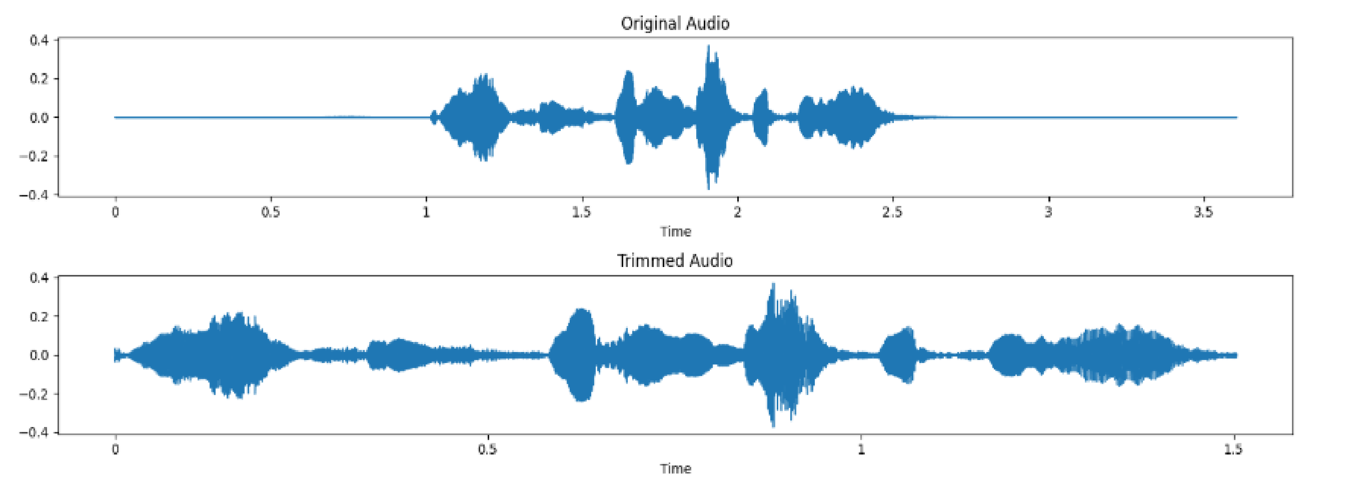
Explanatory Data Analysis

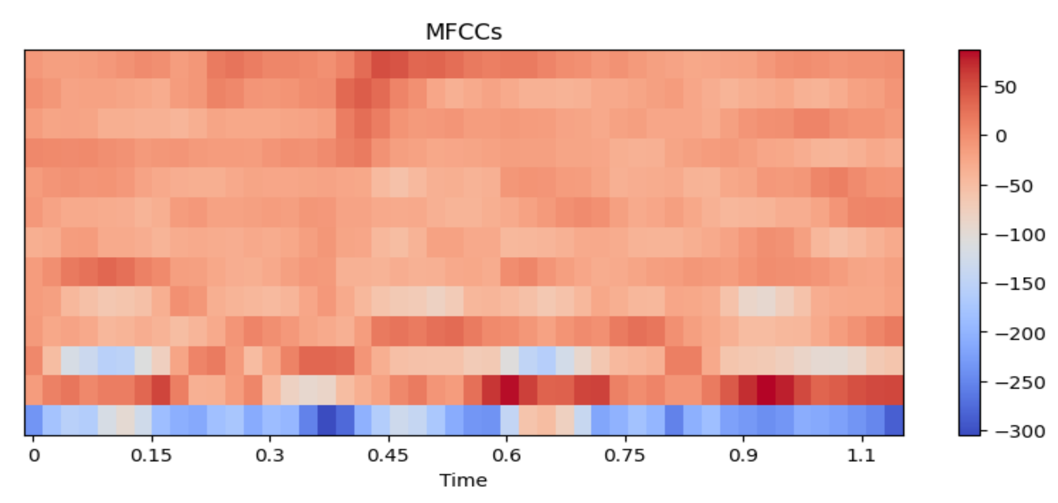
- Speech Signal Visualization
- Frequency Distributions.
- Correlation Analysis.

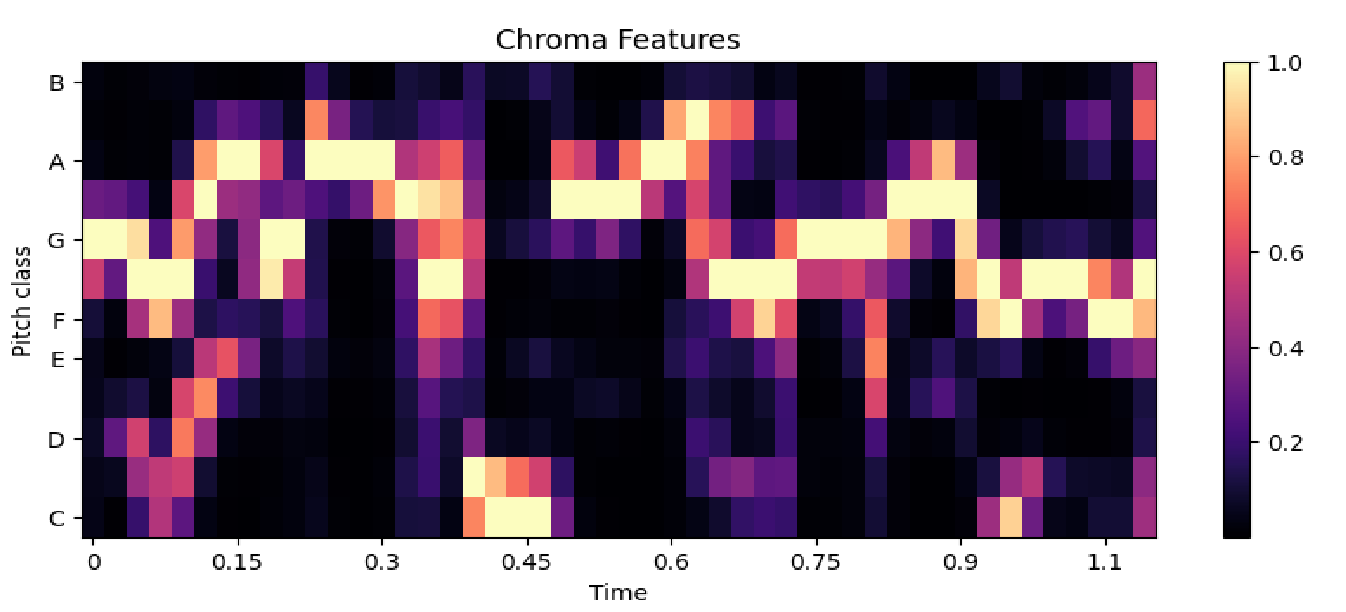


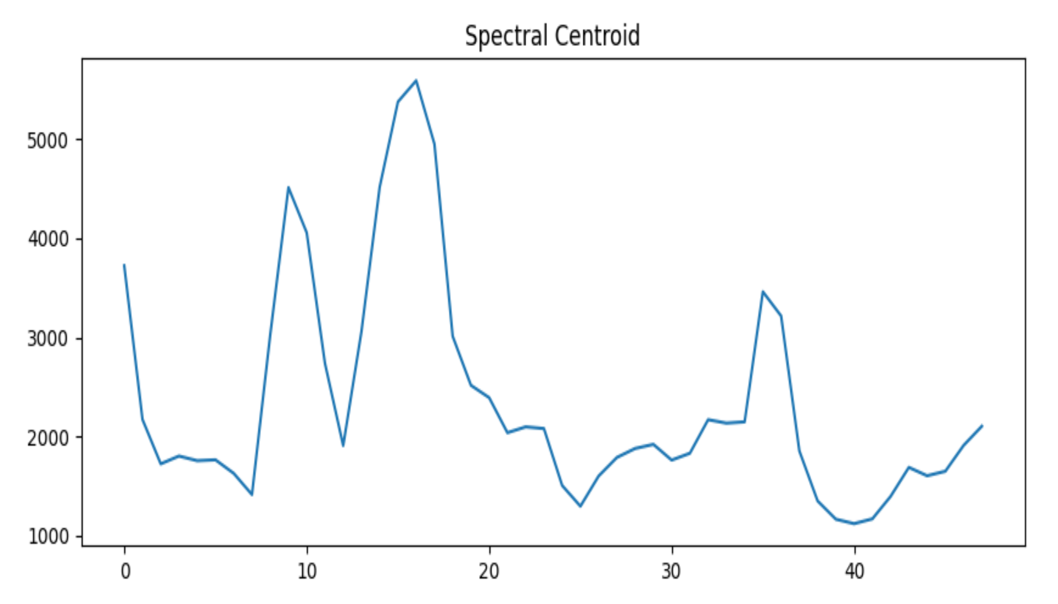


Explanatory Data Analysis









Machine Learning Models Explored

- **Support Vector Maching (SVM)** – useful for high-dimensional spaces in feature extraction.
- **Random Forest** – these help in providing important features and are robust to noise.
- **K-Nearest Neighbors (KNN)** – this is a simple algorithm but has a drawback as computationally expensive in the spaces of high dimensions.
- **Logistic Regression** – mainly useful in binary classification.
- **Multi-Layer Perceptron Classifier (MLP)** – useful in capturing complex, non-linear relationships in the data.

Support Vector Machine (SVM)

- **Why SVM?**
  - ❖It suitably works well for datasets that are of small to medium size.
  - ❖Additionally, it can handle data that is non-linear separable that can be done with kernel methods like RBF, Polynomial.
- **Implementation:**
  - ❖I have used Scikit-learn's SVM model with utilizing hyperparameter tuning features like C, gamma.
- **Results:**
  - ❖I achieved an accuracy of 70.23%

Random Forest

- **Why Random Forest?**
  - ❖One important feature is it reduces the overfitting by doing an average of multiple decision trees.

- **Hyperparameter tuning:**
  - ❖So, the utilized optimal features are n_estimators, max_depth, and min_samples_split.

- **Results:**
  - ❖I achieved an accuracy of 68.45%

K-Nearest Neighbors (KNN)

- **Why KNN?**
  - ❖Well, it is generally simple to implement and is considered as a non-parametric algorithm.

- **Challenges:**
  - ❖So, in this case the performance gradually decreases with high-dimensional data.
- **Results:**
  - ❖I achieved an accuracy of 55.35%

Logistic Regression

- **Why Logistic Regression?**
  - ❖Helps in binary classification(actual presence or absence of speech in the audio)
  - ❖Utilizes the sigmoid function that is used for mapping the outputs between 0 and 1.
  - ❖Easy to interpret like understanding how MFCC and chroma influence predictions.
  - ❖Turns out to be working well for feature extraction.

- **Challenges:**
  - ❖So, this algorithm is not useful for complex relationships – especially when it experiences the data having non-linear patterns.
- **Results:**
  - ❖I achieved an accuracy of 70.238%

Multi-Layer Perceptron Classifier(MLP)

- **Why MLP?**
  - ❖Helps in particularly capturing complex, non-linear relationships in the data overcoming the drawback of linear regression algorithm.
  - ❖This can effectively learn suitable hierarchical representations of any given audio features for example – MFCC and Chroma.
- **Challenges:**
  - ❖So, one of the drawback of this algorithm is it requires intensive tuning of hyperparameters l
- **Results:**
  - ❖I achieved an accuracy of 75.35%

Model Training & Validation

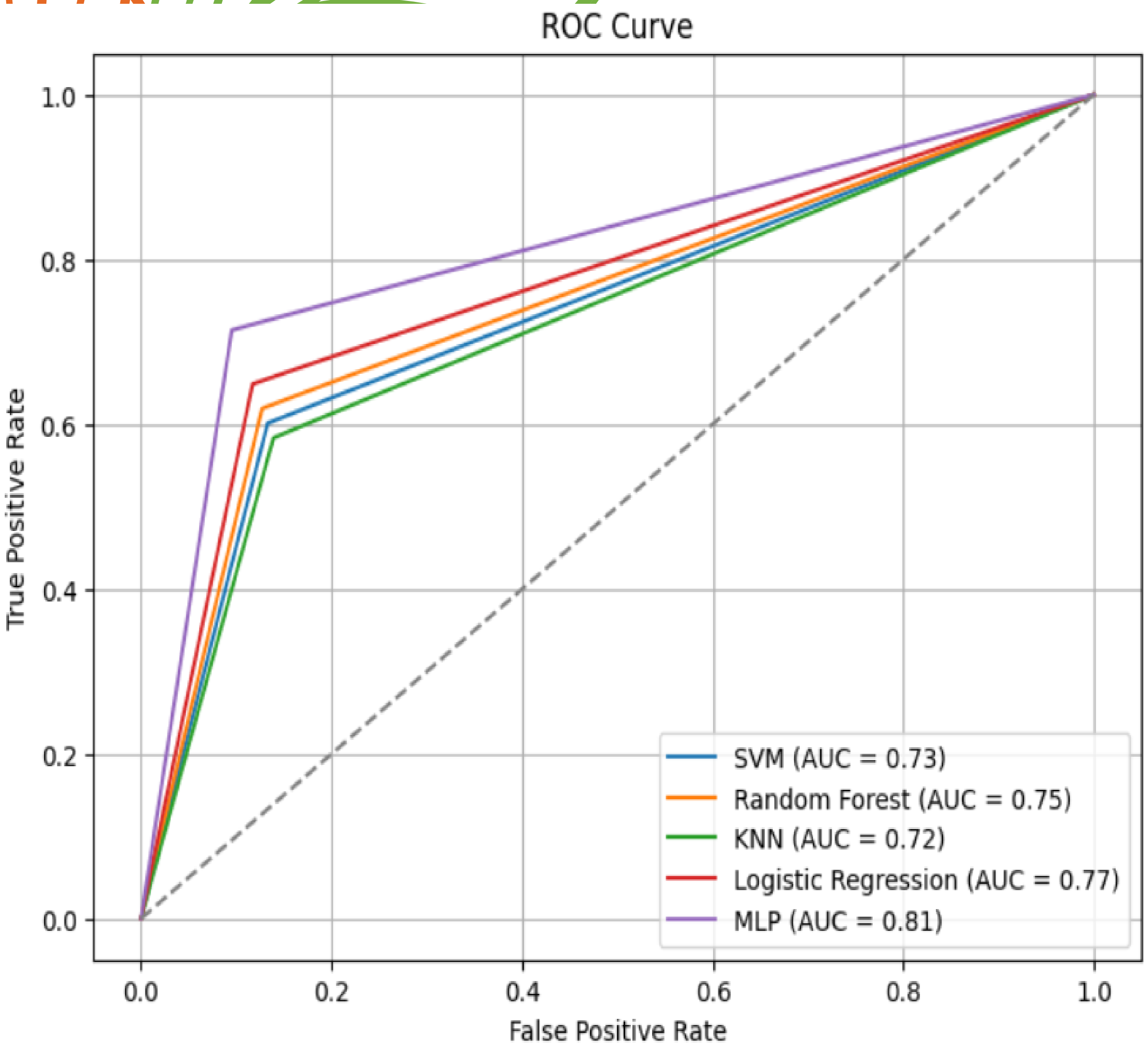
- **Training Process:**
  - ❖Train-Test Split – 75% training, 25% testing.
  - ❖Cross-validation – 5-folds inorder to avoid overfitting.
- **Hyperparameter Tuning:**
  - ❖In here, utilized GridSearchCV for finding the best parameters.

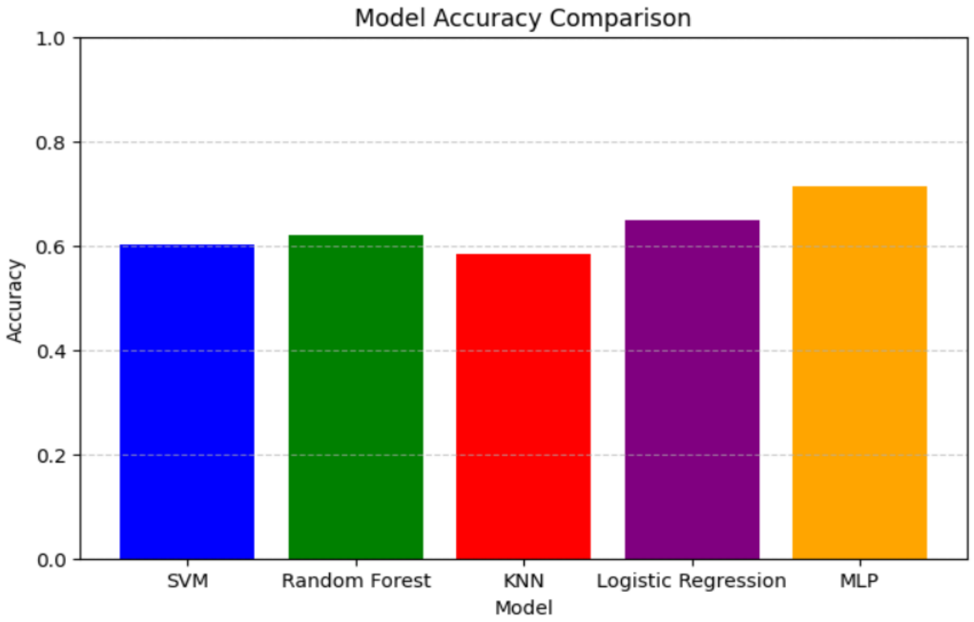
Results & Accuracy Comparison

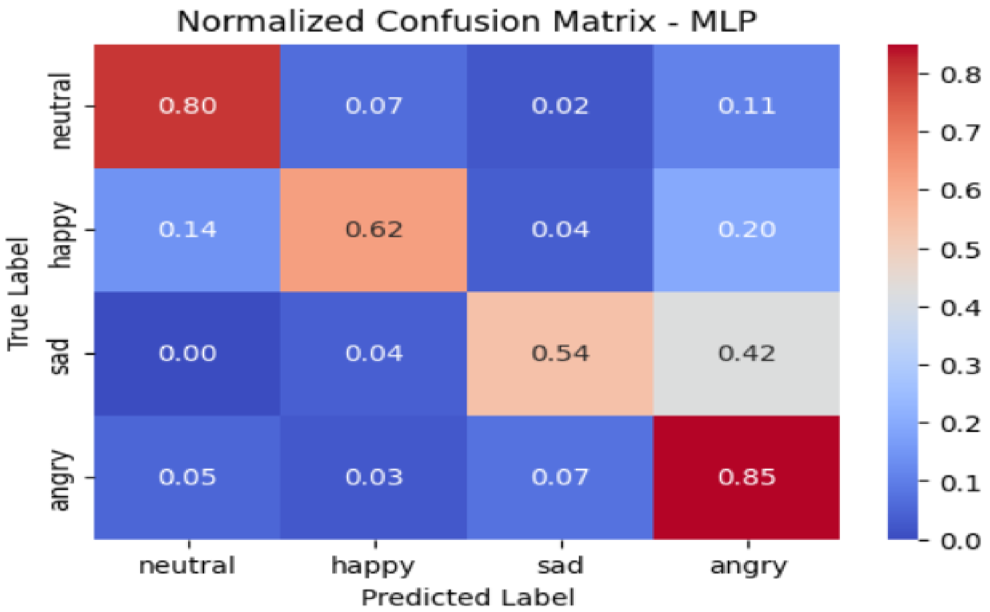
Model	Accuracy (%)
SVM	**68.23%**
Random Forest	**64.45%**
KNN	**55.35%**
Logistic Regression	**70.238%**
Multi Layer Perceptron Classifier	**75.35%**

- **Multi Layer Perceptron classifier** performed the best, followed by **Linear Regression and SVM.**
- While **KNN** algorithm found difficulties with identifying high-dimensional features.

Results & Accuracy Comparison







Challenges Faced

- **Background Noise:**
  - ❖The need for involvement of filtering methods for overlapping voices, echoes, and ambient noise as these factors lower the model performance.
  - ❖So used MFCCs and Log-Mel Spectrograms.
- **Computational Cost:**
  - ❖Algorithms like SVM and KNN are significantly considered to be expensive in computation with the use of MFCCs and chroma features.
- **Data Imbalance:**
  - ❖There are some speech motion classes that have minor samples in the overall dataset, thus leading to biased model predictions.
  - ❖Solution – Implemented Synthetic Minority Over-sampling Technique (SMOTE) to have a synthetic representation of samples for various underrepresented classes.

Limitations

- So, traditional machine learning models are less effective due to temporal dependencies – so in this case we can think of neural networks like RNNs can perform better.
- In terms of dataset biases these impact the overall generalization across different accents/ languages.
- As RAVDESS dataset contains only the audio files in English accent.
- The process of feature extraction is challenging for the real-time processing.

Future Directions

- **Deep Learning Approaches:**
  - ❖In order to elevate the model effectiveness, we can use CNNs, RNNs, or Transformers to perform feature extraction & classification.
- **Real-Time Processing:**
  - ❖To make the system more robust to real time environment, we can optimize the model for low-latency applications.
- **Multi-Modal Learning:**
  - ❖Along with speech we can integrate with video for more richer insights.

Conclusion

- The conventional machine learning techniques offer effective and interpretable results.
- For speech categorization, feature extraction (MFCC, Spectrograms) is essential.
- Deep learning developments in the future may enhance recognition capabilities.
- Further action: Examine hybrid models for real-time speech motion recognition that combine deep learning & conventional machine learning.