

Comparative Study of Machine Learning and Deep Learning Models for Bengali Speech Sentiment Analysis

submitted by

Joydeep Das

Class Roll No: 002210503012

Exam Roll: MCA244012

Master of Computer Application

under the supervision of

Prof. (Dr.) Ujjwal Maulik

in the

Department of Computer Science and Engineering

Faculty of Engineering and Technology

Jadavpur University

2024

Faculty of Engineering and Technology
Jadavpur University

CERTIFICATE OF RECOMMENDATION

To Whom It May Concern

I hereby recommend that the project report entitled “Comparative Study of Machine Learning and Deep Learning Models for Bengali Speech Sentiment Analysis” has been carried out by Joydeep Das (Reg. No.: 163604 of 2022-23, Roll No.: 002210503012, Exam Roll: MCA244012), under my guidance and supervision and be accepted in partial fulfilment of the requirement for the degree of Master of COMPUTER APPLICATION in DEPARTMENT of COMPUTER SCIENCE and ENGINEERING, JADAVPUR UNIVERSITY during the academic year 2023-2024.

Prof. (Dr.) Ujjwal Maulik
Project Supervisor
Dept. of Computer Science and Engineering
Jadavpur University

Prof. (Dr.) Nandini Mukhopadhyay
Head of the Department
Dept. of Computer Science & Engineering
Jadavpur University

Prof. Dipak Laha
Dean, Faculty Council
of Engineering & Technology
Jadavpur University, Kolkata-032

Faculty of Engineering and Technology
Jadavpur University

CERTIFICATE OF APPROVAL

This is to certify that the project report entitled “**Comparative Study of Machine Learning and Deep Learning Models for Bengali Speech Sentiment Analysis**” is a bonafide record of work carried out by JOYDEEP DAS, in fulfillment of the requirements for the award of the degree of Master of Computer Application in the Department of Computer Science and Engineering, Jadavpur University. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the project report only for the purpose for which it has been submitted.

Signature of Examiner 1

Date:

Signature of Examiner 2

Date:

Faculty of Engineering and Technology
Jadavpur University

**DECLARATION OF COMPLIANCE WITH
ACADEMIC ETHICS**

I hereby declare that this project report entitled "Comparative Study of Machine Learning and Deep Learning Models for Bengali Speech Sentiment Analysis" is an outcome of my genuine and independent work, as part of my Master of Computer Application program.

All information in this document has been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, all materials and results that are not original to this work have been fully cited and referenced.

Joydeep Das

Registration No.: 163604 of 2022-23

Examination Roll No.: MCA244012

Thesis Title: - **Comparative Study of Machine Learning and Deep Learning Models for Bengali Speech Sentiment Analysis**

Signature with Date

ACKNOWLEDGEMENT

With my most sincere and gratitude, I would like to thank Prof. (Dr.) Ujjwal Maulik, Department of Computer Science & Engineering, my supervisor, for his overwhelming support throughout the duration of the project. His motivation always gave me the required inputs and momentum to continue with my work, without which the project work would not have taken its current shape. His valuable suggestions and numerous discussions have always inspired new ways of thinking. I feel deeply honored that I got this opportunity to work under him.

I would like to express my sincere thanks to all my teachers for providing sound knowledge base and cooperation.

I would like to thank all the faculty members of the Department of Computer Science & Engineering of Jadavpur University for their continuous support.

Last, but not the least, I would like to thank my batch mates for staying by my side when I needed them.

Joydeep Das
Class Roll No.: 002210503012
Examination Roll No.: MCA244012
Department of Computer Science & Engineering
Jadavpur University

CONTENTS

Abstract	6
1. Introduction	7 - 15
1.1 Machine Learning	8 - 10
1.1.1 Supervised Learning	8 - 9
1.1.2 Unsupervised Learning	10
1.2 Deep Learning	10 - 11
1.3 Sentiment Analysis	11 - 12
1.4 Motivation for Bengali Speech Sentiment Analysis	12
1.5 Audio Features	13
1.6 Related Work on Speech Sentiment Analysis	14 - 15
2. Study of Bengali Speech Sentiment Analysis using Mel-Frequency Cepstral Coefficients Feature	16 - 36
2.1 Dataset Overview	17 - 19
2.1.1 SUBESCO	17 - 18
2.1.2 BanglaSER	19
2.2 Methodology	20 - 28
2.2.1 Feature Description	20 - 21
2.2.2 Data Augmentation	21
2.2.3 Data Preprocessing	22
2.2.4 Machine Learning Models	22 - 25
2.2.5 Deep Learning Models	25 - 27
2.3 Experimental Result	28 - 36
2.3.1 Experimental Setup and Steps	28 - 30
2.3.2 Results	30 - 36
2.4 Discussion	36
3. Study of Bengali Speech Sentiment Analysis using Mel-Frequency Cepstral Coefficients Feature and Chroma Shift Feature	37 - 45
3.1 Methodology	38
3.1.1 Feature Description	38
3.1.2 Machine Learning and Deep Learning Models	38
3.2 Experimental Result	39 - 45
3.2.1 Experimental Setup	39
3.2.2 Results	39 - 45
3.3 Discussion	45
4. Conclusion and Future Work	46 - 48
Reference	49 - 51

Abstract -

Sentiment analysis, a key area within Natural Language Processing (NLP), has garnered significant research attention across various languages. With the rise of Artificial Intelligence (AI), the use of speech in multiple sectors is expanding, highlighting the importance of Speech based Sentiment Analysis. Understanding emotions is crucial for many applications, as they form a fundamental component of human communication. Predicting human emotion from speech is a promising research topic with substantial potential for advancements in Human-Computer Interaction (HCI).

Despite Bengali being the seventh most spoken language globally, research on Bengali speech sentiment analysis remains limited. In this study, we conducted a comparative analysis of sentiment analysis for the Bengali language using both machine learning and deep learning models. Specifically, we evaluated Random Forest Classifier (RFC), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP) algorithms alongside Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) models.

We utilized the SUBESCO and BanglaSER datasets to train our models. Our focus was on recognizing five emotional states: joy, sadness, anger, surprise and neutral. Initially, we employed Mel-Frequency Cepstral Coefficients (MFCC) feature extraction in Chapter 1, and later incorporated Mel-Frequency Cepstral Coefficients (MFCC) and Chroma Shift feature extraction methods in Chapter 2 to offer a comprehensive comparison between the models.

Our experimental evaluations demonstrated the effectiveness of these algorithms in accurately predicting emotions in Bengali speech samples. The results indicated that the accuracy rates ranged mostly around 85% to 95%, with Multi-Layer Perceptron achieving the highest. The Majority Voting model, which combined the results from all models to enhance prediction accuracy, also performed effectively.

The insights gained from this study provide valuable inputs for future explorations and improvements in the domain of Bengali speech sentiment analysis.

CHAPTER 1

INTRODUCTION

1. Introduction

This chapter introduces Machine Learning, covering its various types with examples, delves into Deep Learning with an example, discusses Sentiment Analysis, outlines the motivation for this project, explains Audio features, and reviews related work relevant to this project.

1.1. Machine Learning

Machine learning [1] [2], a subfield of artificial intelligence, represents a transformative approach to data analysis and pattern recognition, enabling systems to learn and improve from experience without being explicitly programmed. This paradigm leverages algorithms and statistical models to analyze and interpret complex datasets, facilitating the development of predictive models and automated decision-making processes. As the volume of data generated by digital activities continues to grow exponentially, machine learning has become indispensable in a multitude of domains, from diagnosing healthcare conditions and predicting financial trends to navigating autonomous vehicles, processing natural language and many more. By continuously refining its capabilities through iterative processes, machine learning stands at the forefront of technological advancement, driving innovation and efficiency across various sectors.

1.1.1. Supervised Learning

In supervised learning [3] [4], algorithms learn from labeled training data, where each input is paired with a corresponding output. The model's objective is to learn a mapping from inputs to outputs and make accurate predictions for new, unseen data. Common applications include classification tasks, such as spam detection and image recognition, and regression tasks, such as predicting house prices or stock values. Popular algorithms in this category include linear regression, logistic regression, support vector machines (SVM), and neural networks.

Example -

- a) Linear regression [5], serves as a fundamental framework for forecasting a continuous output variable by leveraging one or more input features. Its premise lies in assuming a linear correlation between the input variables and the resultant output. It is widely used in applications like predicting house prices and economic forecasting.

- b) Logistic Regression [6], is used for binary classification tasks rather than regression. It models the probability of a binary outcome using a logistic function, producing outputs between 0 and 1. The model is effective for problems such as email spam detection, disease diagnosis, and credit scoring.
- c) Support Vector Machines (SVM) [7], is a versatile model used for both classification and regression tasks. For classification, it identifies the optimal hyperplane that maximizes the margin between data points of different classes.
- d) Decision Trees [8], are intuitive models that split the data into subsets based on the values of input features, forming a tree-like structure. Each node signifies a decision based on a feature, while each branch denotes an outcome. Decision trees are easy to interpret and can handle both classification and regression tasks. They are used in various fields, including finance, healthcare, and marketing.
- e) Random forest [9], is an ensemble learning method that builds multiple decision trees and merges their predictions to improve accuracy and robustness. Each tree in the forest is trained on a different subset of the data, and the final prediction is made by averaging the predictions of all trees (for regression) or by majority voting (for classification).
- f) K-Nearest Neighbors (KNN) [10], is a simple, instance-based learning algorithm used for classification and regression. It makes predictions by identifying the 'k' most similar instances (neighbors) in the training data and using their values to predict the output for new instances. Despite its simplicity, KNN is effective in many practical applications, such as image recognition and recommendation systems, but it can be computationally intensive for large datasets.
- g) Neural Networks [11], are inspired by the human brain's structure and are capable of modeling complex relationships in data. They consist of layers of interconnected nodes (neurons) that process input data and learn features through backpropagation. Neural networks are particularly powerful for tasks involving large datasets and intricate patterns, such as image and speech recognition, natural language processing, and time series forecasting.

1.1.2. Unsupervised Learning

Unsupervised learning [12] deals with unlabeled data, where the algorithm's goal is to uncover hidden patterns or intrinsic structures within the data. This type of learning is particularly useful for clustering, dimensionality reduction, and anomaly detection. Clustering algorithms, like k-means and hierarchical clustering, group similar data points together, while dimensionality reduction techniques, such as principal component analysis (PCA) reduce the number of features in the dataset while preserving important information. Unsupervised learning is widely used in market segmentation, recommendation systems, and biological data analysis.

Example -

- a) K-Means Clustering [13], is a popular clustering algorithm that partitions data into 'k' clusters based on feature similarity. It iteratively assigns data points to clusters by minimizing the sum of squared distances between points and the cluster centroids. K-means is computationally efficient and is commonly used for market segmentation, customer profiling, and image compression.
- b) Hierarchical Clustering [14], builds a tree-like structure (dendrogram) to represent data relationships. It can be either agglomerative (bottom-up approach) or divisive (top-down approach). Agglomerative clustering starts with each data point as a single cluster and merges them iteratively, while divisive clustering starts with all points in one cluster and splits them iteratively. This method is useful for gene expression analysis, social network analysis, and document clustering.
- c) Principal Component Analysis (PCA) [15], is a dimensionality reduction technique that transforms data into a new coordinate system where the greatest variances are captured in the first few principal components. PCA is widely used in image processing, finance, and genomic data analysis.

1.2. Deep Learning

Deep learning [16] [17], a subset of machine learning, focuses on algorithms inspired by the structure and function of the brain's neural networks. Characterized by its use of multiple layers of artificial neurons, deep learning excels at automatically discovering

intricate patterns and representations in large datasets. This advanced technique has revolutionized various fields, including computer vision, natural language processing, and speech recognition, by enabling significant breakthroughs in tasks such as image classification, language translation, and voice assistants. Deep learning's ability to model complex, non-linear relationships in data has made it an indispensable tool in modern artificial intelligence, driving innovation and improving performance across a wide array of applications.

Example -

- a) Convolutional Neural Networks (CNNs) [18] [19], are designed primarily for processing structured grid data like images. They consist of convolutional layers that apply filters to input data, capturing spatial hierarchies and features such as edges, textures, and shapes. CNNs have achieved remarkable success in image classification, object detection, and medical image analysis.
- b) Recurrent Neural Networks (RNNs) [20] [21], are designed to handle sequential data by maintaining a hidden state that captures information from previous time steps. This makes them suitable for tasks involving temporal dependencies, such as time series forecasting, speech recognition, and language modeling.
- c) Generative Adversarial Networks (GANs) [22], consist of two neural networks: a generator and a discriminator. The generator creates synthetic data samples, while the discriminator evaluates their authenticity against real data. GANs have revolutionized fields like image synthesis, style transfer, and data augmentation.

1.3. Sentiment Analysis

Sentiment analysis [23] [24] [25], also known as opinion mining, is a natural language processing (NLP) technique that aims to determine the emotional tone behind a body of text. By analyzing subjective information, sentiment analysis helps identify and categorize opinions expressed in various forms of communication, such as reviews, social media posts, and articles. This powerful tool is widely used in fields like marketing, customer service, and social sciences to gauge public opinion, enhance user experience, and monitor brand reputation.

In addition to text-based analysis [26] [27], sentiment analysis has expanded into the realm of speech, known as speech-based sentiment analysis. This approach leverages audio data to assess emotions conveyed through vocal features such as tone, pitch, and rhythm. Speech-based sentiment analysis is particularly valuable in real-time applications like customer service call centers, where understanding the customer's emotional state can lead to more empathetic and effective responses. By combining text and speech-based sentiment analysis, researchers and practitioners can achieve a more comprehensive understanding of human sentiment across different communication channels.

1.4. Motivation for Bengali Speech Sentiment Analysis

In recent decades, sentiment analysis has predominantly focused on textual data through text mining techniques, leaving a substantial gap in the analysis of audio data. This gap is particularly significant for under-resourced languages like Bengali. Sentiment analysis involves identifying the sentiment polarity of a piece of text, categorizing it as positive, negative, or neutral based on contextual cues.

Despite the rich emotional information embedded in audio data, research in audio speech analysis has lagged behind textual analysis. This is noteworthy given that audio data can capture a broader range of human emotions, providing unique insights into sentiment. For languages like Bengali, which is spoken by approximately 300 million people worldwide, the scarcity of resources has led to limited research in sentiment analysis. The field remains largely unexplored in the context of Bengali audio data.

This project aims to address this gap by enhancing communication technologies and developing applications that can accurately detect, interpret, and respond to human emotions in Bengali. By focusing on audio data, this project seeks to deepen our understanding of emotional expression in Bengali speech, ultimately improving sentiment analysis capabilities for the Bengali-speaking population.

1.5. Audio Features

Audio features [28] [29] are distinctive attributes or characteristics extracted from audio signals, used to analyze and process sound in various applications, including speech recognition, music classification, and sentiment analysis. These features capture essential information about the audio signal, allowing for efficient and effective analysis.

There are several categories of audio features, each providing different insights into the audio content:

- **Time-Domain Features:** These features are derived directly from the audio waveform and include characteristics such as zero-crossing rate, which measures the rate at which the signal changes sign, and energy, which represents the signal's power over time.
- **Frequency-Domain Features:** These features are obtained by transforming the audio signal into the frequency domain using techniques like the Fast Fourier Transform (FFT). Examples include spectral centroid, which indicates the "center of mass" of the spectrum, and spectral bandwidth, which measures the range of frequencies present in the signal.
- **Cepstral Features:** These are derived from the spectrum of the logarithm of the signal's spectrum and include Mel-Frequency Cepstral Coefficients (MFCCs). MFCCs are particularly effective in capturing perceptually relevant aspects of audio signals, making them widely used in speech and audio processing.
- **Rhythmic Features:** These features capture the temporal patterns and rhythms in audio signals, often used in music analysis. Examples include tempo and beat strength.
- **Harmonic Features:** These features describe the harmonic content of the audio signal, including pitch and harmonic ratios, which are crucial for tasks like music genre classification and speaker identification.

By extracting and analyzing these features, researchers and engineers can develop algorithms and systems that effectively interpret and manipulate audio data, enabling advancements in various audio-related technologies.

1.6. Related Work on Speech Sentiment Analysis

The field of emotion recognition from speech has seen a significant evolution, employing various methods ranging from traditional prosodic analysis to advanced machine learning (ML) and deep learning (DL) techniques. This survey organizes the related work into three main categories: studies without machine learning, those utilizing machine learning, and those incorporating deep learning.

Research on sentiment and emotion analysis from speech has historically relied on prosodic and acoustic features without the use of advanced machine learning techniques. For instance, Luengo et al. [30] utilized prosodic parameters to automatically identify emotions in Basque speech using Gaussian Mixture Models (GMM) and Support Vector Machines (SVM). They achieved up to 98.4% accuracy with spectral features and GMM, demonstrating the utility of prosodic information in emotion detection.

1.6.1. Machine Learning Based Approach

Machine learning approaches have significantly advanced the field by leveraging powerful algorithms to classify emotions based on extracted speech features such as Mel-Frequency Cepstral Coefficients (MFCC).

Devnath et al. [31] focused on recognizing emotions such as happiness, sadness, anger, and neutrality using a k-Nearest Neighbor (k-NN) classifier. Their self-built dataset showed an average accuracy of 87.50%, with individual emotion accuracies ranging between 75% and 85%.

Prashengit Dhar and Sunanda Guha [32] proposed a system for predicting emotions from Bengali speech using MFCC and Linear Predictive Coding (LPC) features. They applied various machine learning algorithms and found that Logistic Regression performed best in recognizing emotions like anger, happiness, and neutrality from the self-collected ABEG dataset and the popular RAVDESS dataset.

Islam et al. [33] explored the use of Random Forest, XGBoost, CatBoost, and Gradient Boosting for emotion classification in Bengali speech. Using MFCC features, CatBoost achieved the highest accuracy of 82.85%, highlighting the effectiveness of these machine learning models in emotion recognition tasks.

Sultana and Rahman [34] conducted an analysis and optimization of acoustic features for recognizing emotions in Bangla speech. They utilized SVM, Random Forest, and XGBoost algorithms, achieving competitive accuracy with optimized features through Grid Search

and five-fold cross-validation, suggesting that machine learning models can perform comparably to deep learning models with lower computational complexity.

1.6.2. Deep Learning Based Approach

Deep learning models have further pushed the boundaries of speech emotion recognition by capturing complex patterns in audio data through sophisticated architectures like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks.

Hasan and Islam [35] used Recurrent Neural Networks (RNN) for emotion recognition in Bengali speech. They employed MFCC and Modulation Spectral (MS) features, achieving 51.33% accuracy with a novel Bengali speech dataset when using RNN modulation-based categorization.

Hasan et al. [36] presented a hybrid approach using Bi-LSTM for Bengali speech emotion recognition, integrating MFCC, Chroma, and Mel-Spectrogram features. Their model showed an overall accuracy of 83.33%, outperforming SVM and Multi-Layer Perceptron (MLP) classifiers in detecting seven different emotion classes.

Sultana et al. [37] proposed a deep learning approach combining Deep Convolutional Neural Networks (DCNN) and Bidirectional LSTM (BLSTM) with a time-distributed flatten (TDF) layer for speech emotion recognition. Their model achieved state-of-the-art performance with weighted accuracies of 86.9% on the SUBESCO dataset and 82.7% on the RAVDESS dataset. This study demonstrated the effectiveness of combining CNN and LSTM architectures for capturing both temporal and sequential emotional cues in speech.

Shruti et al. [38] conducted a comparative study on Bengali speech sentiment analysis using machine learning models (including KNN and AdaBoost) and deep learning models (CNN, LSTM, Bi-LSTM). Their study revealed that the KNN model achieved 90% accuracy on the SUBESCO and BanglaSER datasets, while AdaBoost and Bi-LSTM models performed best with 45% accuracy on a custom test dataset. They also utilized the SHAP model for interpretability of the machine learning model outcomes.

These studies collectively advance the field of Bengali speech sentiment and emotion analysis, demonstrating the progression from traditional prosodic feature analysis to sophisticated machine learning and deep learning models, each contributing uniquely to improve accuracy and understanding in this domain.

CHAPTER 2

STUDY OF BENGALI SPEECH SENTIMENT ANALYSIS USING MEL-FREQUENCY CEPSTRAL COEFFICIENTS FEATURE

This chapter provides an overview of the datasets used, fundamental concepts, methodologies used in this project, and results using mfcc feature extraction, which is one of the most commonly used features to work with audio data.

2.1. Dataset Overview

A dataset is a structured collection of data, often organized in tables, which is used for analysis and model training in various fields such as machine learning, statistics, and data science. The quality and relevance of the dataset are crucial, as they directly impact the performance and accuracy of analytical models and algorithms. By providing a representative sample of real-world data, a well-curated dataset enables researchers and practitioners to develop, validate, and refine predictive models, ensuring they generalize effectively to new, unseen data.

For this project, we have used two datasets to train and test our models. These two datasets are:

2.1.1. SUBESCO

SUST Bangla Emotional speech corpus (SUBESCO) [39] is a public audio-only emotional speech dataset for Bengali language which consists of 7000 audio speech labeling 7 classes of emotions namely: Happy, Sad, Disgust, Fear, Angry, Neutral, and Surprise. The gender-balanced collection included trained native speakers, and each recording of 10 sentences simulated the seven emotions consisting of an accuracy rate of 71%- 80%. The average file length of each audio file is 4s and all the audio files are available in .wav format.

Figure 1 shows the waveplot of emotions in SUBESCO dataset.

Emotion	Count
Happy	1000
Sad	1000
Disgust	1000
Fear	1000
Angry	1000
Neutral	1000
Surprise	1000

Table 1: Number of records of each emotion

Filename in this dataset is 6 fields separated by underscore (_).

Filename Convention is:

Gender (M/F)_Speaker number (01 - 10)_Speaker name_Sentences (S_1 - S_10)_State of emotion_Repetition (1 - 5).

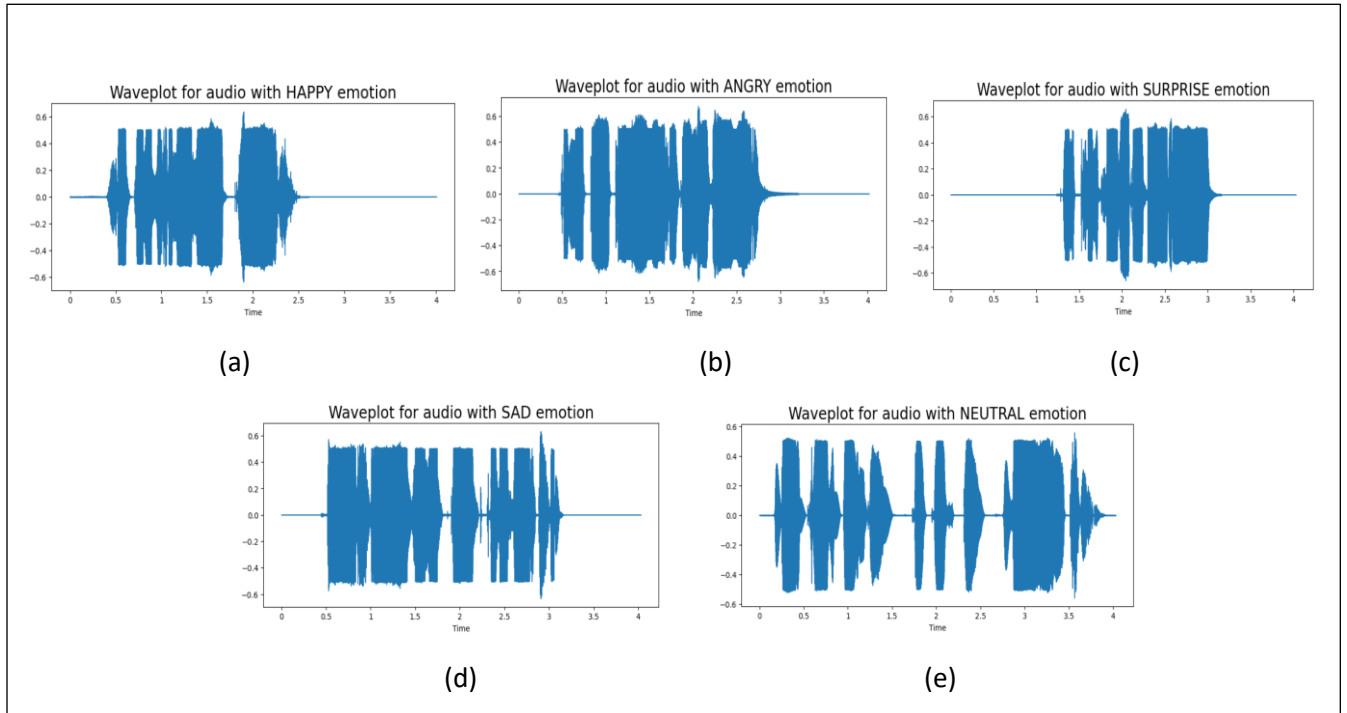


Fig 1: Waveplot of different emotions in SUBESCO dataset

2.1.2. BanglaSER

BanglaSER [40] consists of 1467 Bangla speech-audio recordings made by an equal number of male and female speakers including 5 different emotions namely Happy, Sad, Angry, Surprise, Neutral. All the audio files are available in .wav format.

Figure 2 shows the waveplot of emotions in BanglaSER dataset.

Emotion	Count
Happy	306
Sad	306
Angry	306
Surprise	306
Neutral	243

Table 2: Number of records of each emotion

Filename in this dataset is 7 fields separated by underscore (_).

Filename Convention is: Mode (03)_Statement type (01)_Emotion (01 - 05)_Intensity (01 - 02)_Statements (01 - 03)_Repetition (01 - 03)_Actor (01 - 34).

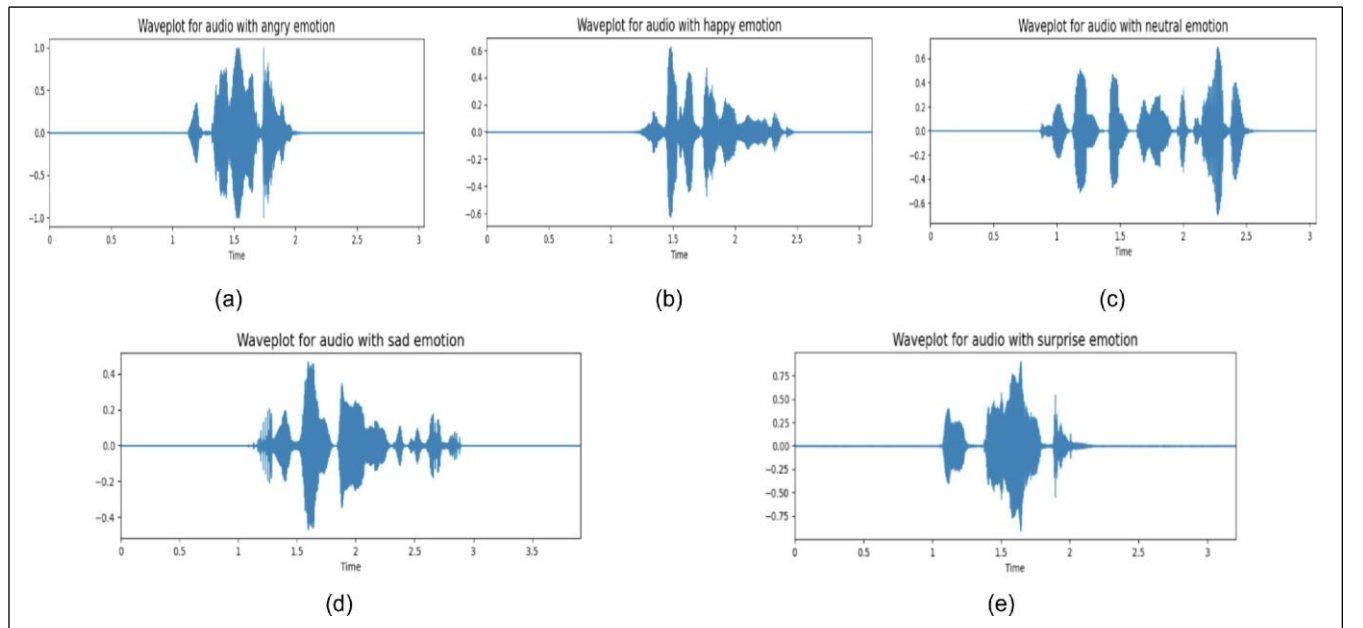


Fig 2: Waveplot of different emotions in BanglaSER dataset

2.2. Methodology

In this section, we have covered the MFCC feature and its working principle, the data augmentation and preprocessing techniques used in the project, as well as the machine learning and deep learning models employed.

2.2.1. Feature Description

Mel-Frequency Cepstral Coefficients (MFCC) [41] [42], is a feature extraction technique widely used in speech and audio processing applications. Introduced in the early 1980s by Davis and Mermelstein, MFCCs have become a fundamental tool for tasks such as speech recognition, speaker identification, and sentiment analysis. The technique aims to capture the important features of audio signals that are most relevant to human perception, making it particularly effective for analyzing speech.

The history of MFCCs dates back to efforts in the 1970s and 1980s to improve the accuracy of speech recognition systems. Researchers recognized that traditional spectral analysis methods, which focused on the linear frequency scale, were not well-aligned with how humans perceive sound. Human auditory perception is better represented by a non-linear scale, known as the Mel scale, which approximates the human ear's response more accurately. This insight led to the development of MFCCs, which incorporate the Mel scale to provide a more perceptually relevant representation of audio signals.

MFCCs are computed by first converting an audio signal into frames and applying the Fast Fourier Transform (FFT) to each frame to obtain the power spectrum. The spectrum is then mapped onto the Mel scale, and the logarithm of the Mel spectrum is taken. Finally, the Discrete Cosine Transform (DCT) is applied to the log Mel spectrum to produce the cepstral coefficients, which form the MFCCs. This process effectively compresses and highlights the most significant features of the audio signal, facilitating more accurate and efficient analysis. Figure 3 picturizes this extraction process.

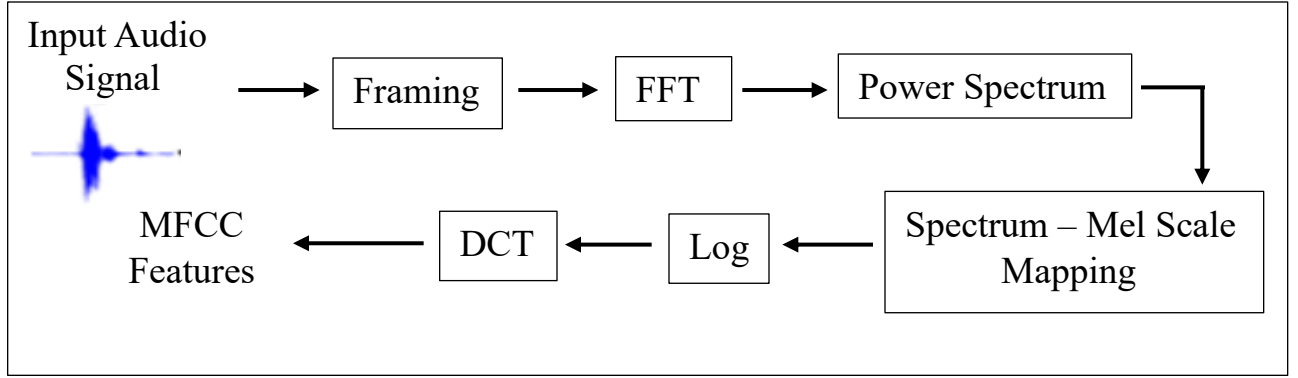


Fig 3: MFCC Feature Extraction

MFCCs are favored for their ability to represent the short-term power spectrum of a sound in a compact form, capturing essential features while discarding redundancies and irrelevant information. This makes them particularly useful in speech-related applications, where they help in distinguishing different phonetic elements and speaker's characteristics effectively. As a result, MFCCs have become a fundamental tool in modern speech and audio processing technologies.

`librosa.feature.mfcc` was used to compute the pandas data frame of our dataset consisting of 40 features (`n_mfcc = 40`).

2.2.2. Data Augmentation

Normally, datasets contain audio recorded in controlled environments, where many parameters are meticulously managed. However, real-life data often contains noise and various imperfections. To improve the generalization capability of our models and make the data more representative of real-world conditions, we employed data augmentation techniques to create syntactic data.

For augmentation, we used the `librosa` library to add noise and stretch to the raw audio in our dataset. First, we injected random noise into the audio to simulate a noisy environment. Next, we applied a stretch factor of 0.8, which slowed down the audio, increasing its duration. These augmented audio samples were then vertically stacked with the original audio data using numpy's `'vstack'` function to enhance the dataset diversity and robustness. This approach helped our models better handle variability and improve performance on real-world noisy data.

2.2.3. Data Preprocessing

Preprocessing data is essential for preparing datasets for machine learning models, ensuring they perform optimally. For this project, we utilized standard techniques to preprocess our data.

The initial preprocessing involved standardizing the feature values using the StandardScaler [43] from scikit-learn, which scales the data to have zero mean and unit variance. This step was applied to both the training and test sets. If the training data was available, it was transformed accordingly; otherwise, only the test data was scaled.

For ML models, this initial preprocessing step was sufficient. However, for DL models, additional preprocessing was required.

In the additional preprocessing stage, the data dimensions were expanded to meet the input shape requirements of DL models. Specifically, the data was reshaped to add an extra dimension.

Furthermore, the target labels were converted to one-hot encoded format using the OneHotEncoder [44] from scikit-learn, transforming categorical labels into a binary matrix representation. This encoding process ensures that each label is represented as a vector with a single high (1) and the rest low (0). This step is essential for classification tasks in deep learning models, enabling them to output probabilities for each class.

Through these preprocessing steps, the data was effectively prepared for training and evaluation across various machine learning and deep learning models, ensuring consistency and compatibility with the specific requirements of each model type.

2.2.4. Machine Learning Models

In this project, we have used Random Forest Classifier (RFC), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP) models.

Random Forest Classifier

The Random Forest Classifier (RFC) is an ensemble learning method that combines multiple decision trees to improve classification accuracy and robustness. This technique leverages the power of multiple decision trees by aggregating their predictions to arrive at a final decision.

Leo Breiman, a prominent statistician, developed the Random Forest algorithm in 2001, as an extension of his earlier work on decision trees and bagging (bootstrap aggregating). His goal was to reduce the variance and overfitting commonly associated with single decision trees. By creating an ensemble of trees, each trained on different subsets of the data, Random Forests were able to achieve significant improvements in predictive performance.

The working principle of RFC is as follows-

- a) Bootstrap Sampling: The process begins with creating multiple subsets of the training data using bootstrap sampling, where each subset is generated by randomly sampling with replacement. This means some samples may appear multiple times in a subset, while others may not appear at all.
- b) Training Multiple Trees: For each bootstrap sample, a decision tree is trained. Unlike standard decision trees, Random Forests introduce an additional layer of randomness by selecting a random subset of features at each split point in the tree. This ensures that the trees are diverse and uncorrelated.
- c) Tree Growth: Each decision tree is grown to its full extent without pruning, meaning each tree becomes a deep learner that captures complex patterns in the data. The randomness in feature selection helps prevent the trees from becoming too similar, which enhances the ensemble's overall performance.
- d) Aggregation of Predictions: Once all the trees are trained, the Random Forest classifier makes predictions by aggregating the outputs of the individual trees. For classification tasks, this typically involves a majority voting scheme where each tree votes for a class label, and the label with the most votes becomes the final prediction.

The Random Forest algorithm's strength lies in its ability to handle a large number of features, manage missing data, and avoid overfitting, making it particularly effective for a wide range of classification tasks. The Random Forest classifier has become a cornerstone of modern machine learning due to its simplicity and ability to build robust, accurate models that generalize well to new data. Figure 5 gives an illustration of the RFC model used in this project.

Support Vector Machine

Support Vector Machine (SVM) is a powerful and versatile supervised learning algorithm used for classification and regression tasks. SVMs were first introduced by Vladimir Vapnik and his colleagues at AT&T Bell Labs in the early 1990s. The key idea behind SVM is to find the optimal hyperplane that best separates the data points of different classes in a high-dimensional space.

The working principle of SVM is as follows -

- a) Linear Separation: For linearly separable data, SVM identifies the hyperplane that maximizes the margin between the closest points (support vectors) of the classes. The broader the margin, the stronger the classifier's generalization capability.
- b) Kernel Trick: For non-linearly separable data, SVM uses the kernel trick to map the input features into a higher-dimensional space where a linear separation is possible. Common kernels include the polynomial kernel, radial basis function (RBF), and sigmoid kernel.
- c) Optimization: SVM solves a convex optimization problem to find the hyperplane that minimizes the classification error while maximizing the margin.

SVMs are recognized for their resilience in high-dimensional spaces and efficiency in scenarios where the dimensionality surpasses the sample size. Their ability to handle both linear and non-linear classification problems and their effectiveness in finding complex decision boundaries and their robustness to overfitting has made them a popular choice in various fields such as bioinformatics, text classification, image recognition and more.

Multi-Layer Perceptron

The Multilayer Perceptron (MLP) is a type of artificial neural network that forms the foundation of many modern deep learning architectures. Its development traces back to the mid-20th century, with significant advancements in the 1980s following the introduction of backpropagation by Rumelhart, Hinton, and Williams.

An MLP is composed of three primary types of layers (Figure 4):

- a) **Input Layer**: Receives the input features.
- b) **Hidden Layers**: One or more layers where each neuron applies a weighted sum of inputs followed by a non-linear activation function, enabling the network to learn complex patterns.
- c) **Output Layer**: Produces the final prediction.

The working principle of MLP is as follows -

- a) Forward Pass: Data is passed through the network layer by layer, with each neuron computing a weighted sum of its inputs and applying an activation function (e.g., ReLU, sigmoid).
- b) Backpropagation: During training, the network adjusts the weights using backpropagation, which calculates the gradient of the loss function with respect to each weight and updates them to minimize the prediction error.

MLPs are the foundation of more advanced neural networks and have been pivotal in the development of deep learning, providing robust solutions for a wide range of applications including image and speech recognition, and natural language processing. Figure 6 provides a depiction of the MLP model utilized in this project.

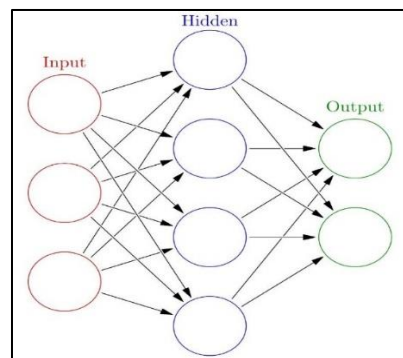


Fig 4: MLP Classifier

2.2.5. Deep Learning Models

In this project, we have used Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNNs) models.

Long Short-Term Memory

Long Short-Term Memory (LSTM) [45] networks are a type of recurrent neural network (RNN) designed to overcome the limitations of traditional RNNs, especially in handling long-term dependencies. Introduced by Hochreiter and Schmidhuber in 1997, LSTMs have become crucial in the fields of sequence prediction and time-series analysis.

The LSTM architecture includes:

- a) **Cell State**: The cell state carries the long-term memory of the network.
- b) **Gates**: LSTMs use three gates to control the flow of information:

- **Forget Gate:** Decides what information to discard from the cell state.
- **Input Gate:** Determines what new information to add to the cell state.
- **Output Gate:** Controls the output based on the cell state and input.

The working of an LSTM involves:

- a) Information Flow: At each time step, the gates manage the cell state by adding or removing information, ensuring important information is retained over long sequences.
- b) Gradient Flow: LSTMs mitigate the vanishing gradient problem, allowing them to maintain gradients over long sequences and thus learn long-term dependencies effectively.

This architecture enables LSTMs to effectively capture long-term dependencies, making them highly effective for tasks involving sequential data, such as time series forecasting, speech recognition, and natural language processing. Figure 7 picturizes the LSTM model used in this study.

Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a class of deep neural networks commonly applied to analyze visual imagery. The development of CNNs can be traced back to the 1980s, but they gained significant attention and advancement in the early 2010s, particularly with the introduction of deep learning techniques.

The working of a CNN involves several key layers:

- a) Convolutional Layers: These layers apply convolution operations to the input, using filters (or kernels) that slide across the input data. This operation captures spatial hierarchies and local patterns, such as edges or textures, within the data. The output of these convolutions is a set of feature maps.
- b) Activation Functions: Typically, non-linear activation functions like ReLU (Rectified Linear Unit) are applied to introduce non-linearity into the model, allowing it to learn more complex patterns.
- c) Pooling Layers: Pooling layers down sample the feature maps by summarizing the presence of features in sub-regions, reducing the spatial dimensions and computational load while retaining essential information.

- d) Fully Connected Layers: After several convolutional and pooling layers, the high-level reasoning in the network is performed via fully connected layers. These layers flatten the output from the previous layers and connect every neuron in one layer to every neuron in the next layer.
- e) Output Layer: For classification tasks, the final layer typically uses a softmax activation function to output probability distributions over the possible classes.

CNNs excel at capturing spatial dependencies and hierarchical patterns, making them highly effective for image and video analysis. Their architecture enables them to learn and generalize features across different parts of the input data, leading to remarkable successes in various domains, including object detection, image segmentation, and even speech and audio processing. Figure 8 provides a depiction of the CNN model utilized in this project.

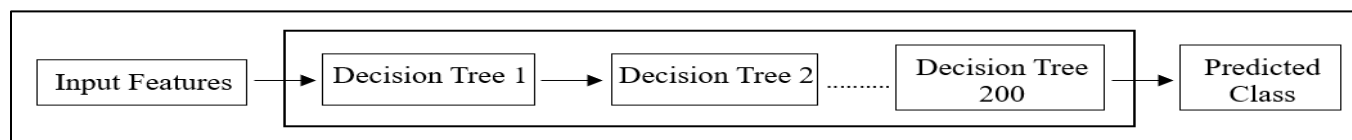


Fig 5: RFC

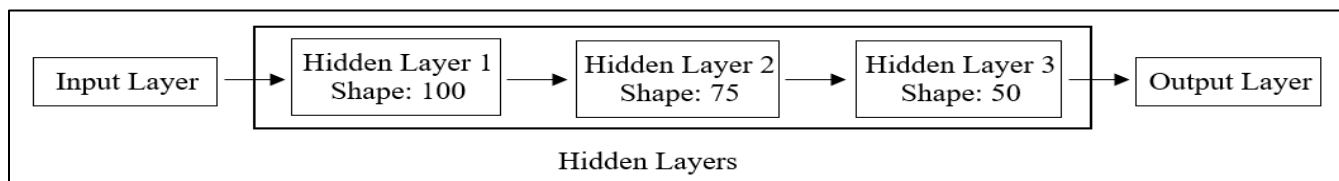


Fig 6: MLP

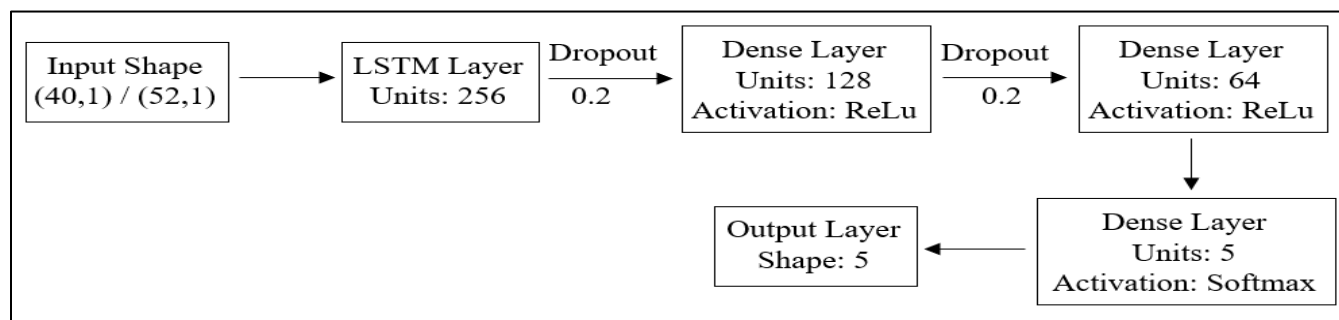


Fig 7: LSTM

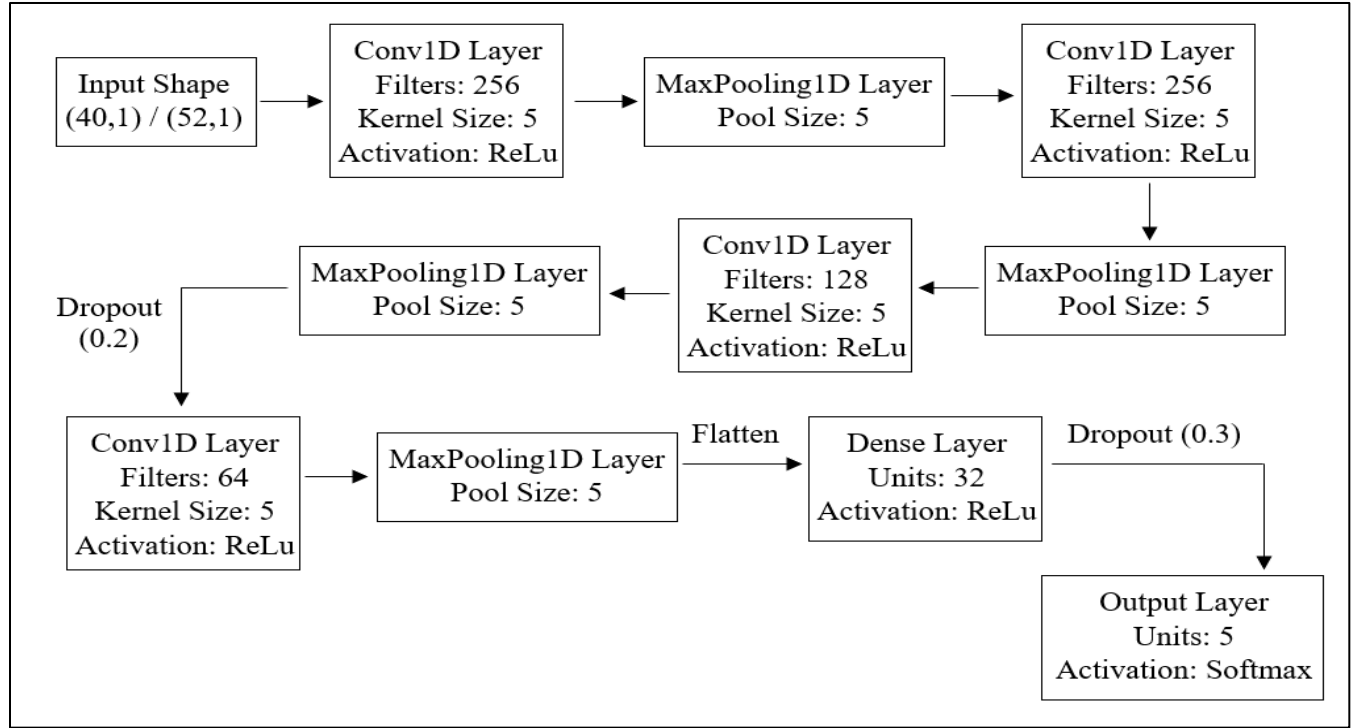


Fig 8: CNN

2.3. Experimental Result

2.3.1. Experimental Setup and Steps

For this project, we utilized Google Colab as the computational environment and employed scikit-learn for implementing the machine learning models. The experimental process was structured into three main steps:

Step 1: From the SUBESCO dataset, we selected five emotions: Happy, Sad, Angry, Surprise, and Neutral, resulting in a total of 5,000 audio samples. The dataset was then divided into an 80:20 ratio, with 80% of the data used for training, out of which 10% is validation data for LSTM and CNN training, and 20% for testing. The training set was utilized to train all the models and the best validation model was saved, while the test set was employed to evaluate the accuracy of the trained models.

Step 2: The BanglaSER dataset was used to test the trained models without any prior training on this dataset. This step aimed to assess the generalization capability of the models on a completely new dataset.

Step 3: The BanglaSER dataset was split into a 60:40 ratio for training (out of which 10% is validation data for LSTM and CNN training) and testing. The training subset was used

to retrain all the models thereby augmenting the models with additional data, and best validation accuracy model was saved. The testing subset was then used to evaluate the performance of the newly trained models.

In both steps 2 and 3, a majority voting ensemble model was implemented. This ensemble approach aggregated the predictions from all individual models and determined the final emotion prediction based on the majority vote. This method aimed to enhance the robustness and accuracy of the emotion prediction system.

Table 4, 5, 6 presents the accuracy result of the above steps respectively.

Figures 9 and 10 display the training curves (accuracy and loss) for LSTM and CNN respectively on SUBESCO dataset.

Figures 11 and 12 display the training curves (accuracy and loss) for LSTM and CNN respectively on BanglaSER dataset.

Figures 13, 14 and 15 illustrate the confusion matrix of the models for the above steps respectively.

Performance Metrics:

To measure the performances of the classification, we utilized two different metrics: Accuracy and Confusion Matrix.

Accuracy: Accuracy is determined by the proportion of correctly predicted instances out of the total instances.

$$\text{Formula: Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

Confusion Matrix: A confusion matrix is a table that describes the performance of a classification model by showing the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Table 3: Confusion Matrix Structure

Training Curve:

Training curves are essential tools in evaluating and understanding the performance of machine learning models during the training process. They typically include two main plots:

Training Loss Curve: This curve illustrates how the loss (error) of the model decreases over the training epochs. The loss is a measure of how well the model's predictions match the actual outcomes. A decreasing loss curve indicates that the model is learning and improving its performance. Ideally, the curve should show a steady decline, converging towards a lower value as training progresses, suggesting that the model is minimizing errors.

Training Accuracy Curve: This curve displays the accuracy of the model on the training dataset over epochs. An increasing accuracy curve signifies that the model is correctly learning to classify the training data. The curve should ideally rise and plateau as the model reaches optimal performance.

Analyzing these curves helps in diagnosing issues such as overfitting, underfitting, and learning rate problems.

2.3.2. Results

Method	Accuracy
RFC	95.27%
SVM	93.93%
MLP	98.37%
LSTM	91.90%
CNN	96.60%

Table 4: SUBESCO dataset (80% training: 20% testing)

Method	Accuracy
RFC	33.79%
SVM	34.60%
MLP	33.33%
LSTM	27.91%
CNN	33.69%
Majority Voting	34.69%

Table 5: BanglaSER dataset (without training)

Method	Accuracy
RFC	85.89%
SVM	79.18%
MLP	86.79%
LSTM	76.92%
CNN	77.03%
Majority Voting	87.13%

Table 6: BanglaSER dataset (60% training: 40% testing)

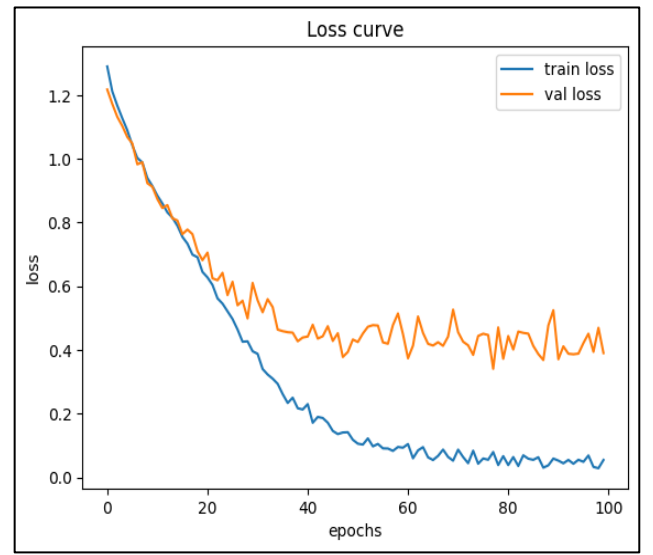
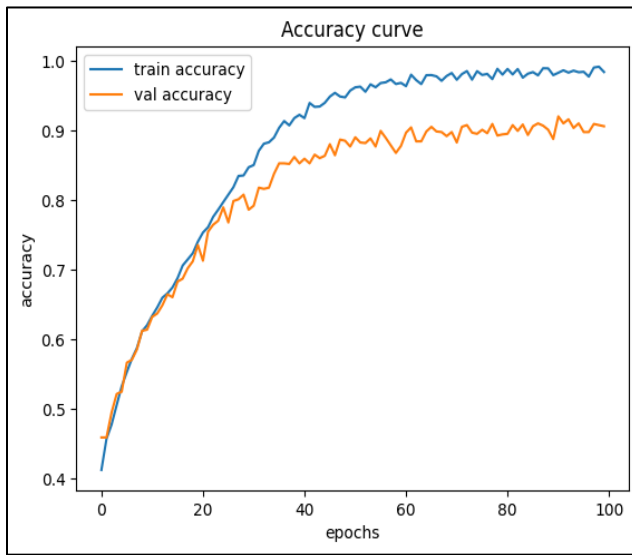


Fig 9: LSTM Training Curves (SUBESCO Dataset)

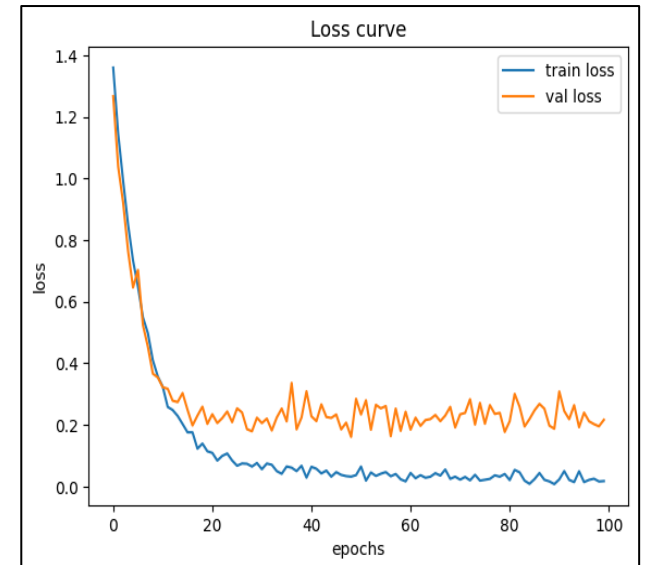
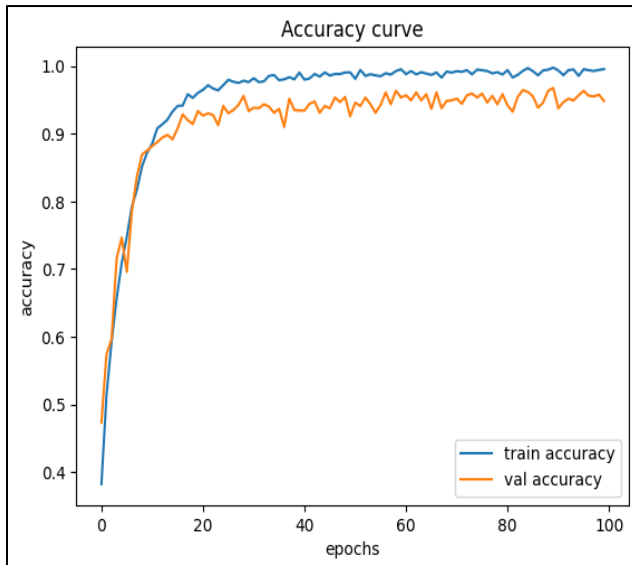


Fig 10: CNN Training Curves (SUBESCO Dataset)

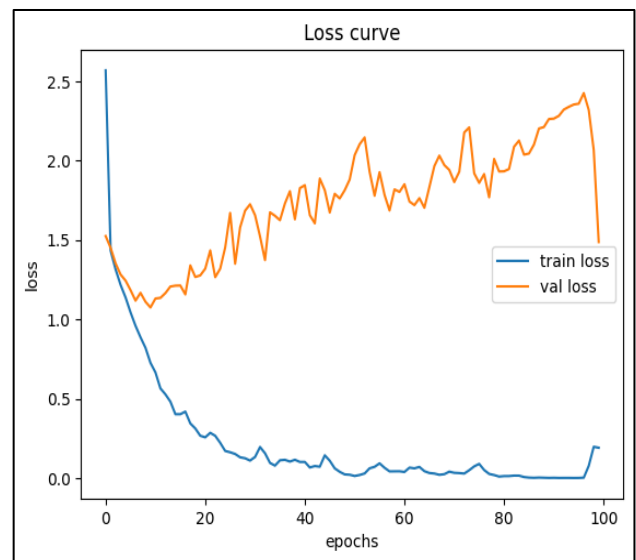
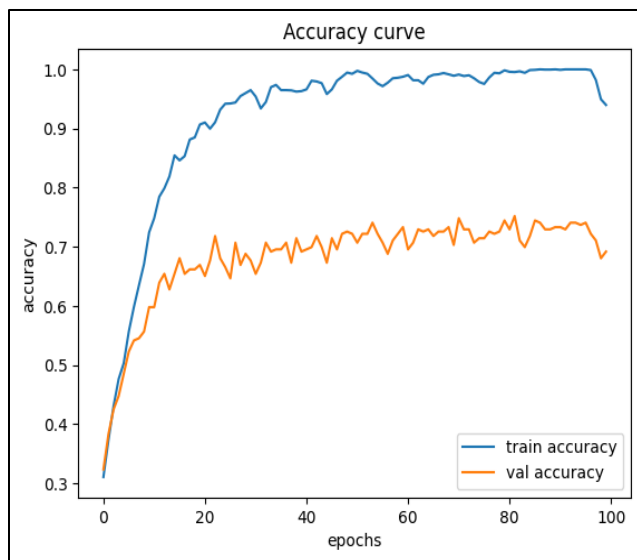


Fig 11: LSTM Training Curves (After Augmenting BanglaSER Dataset)

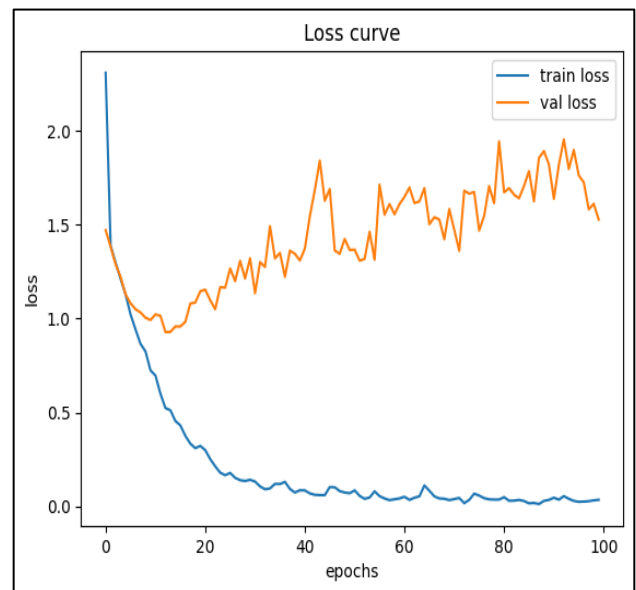
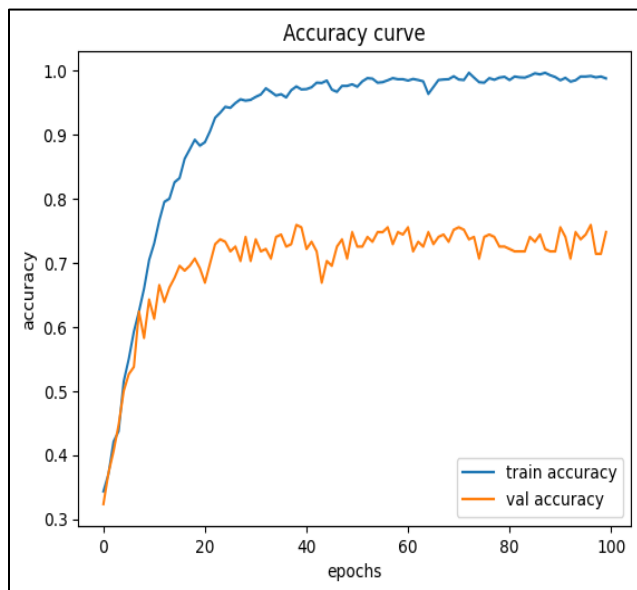
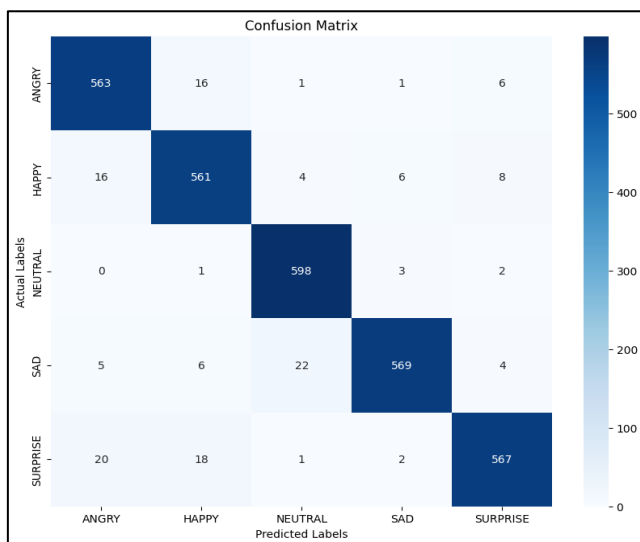
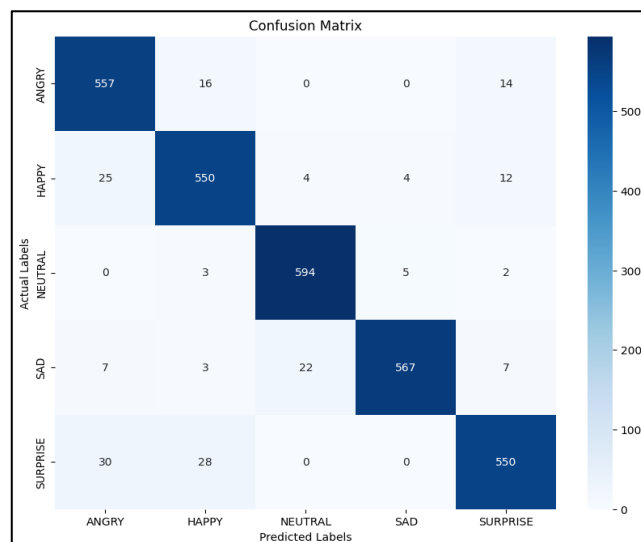


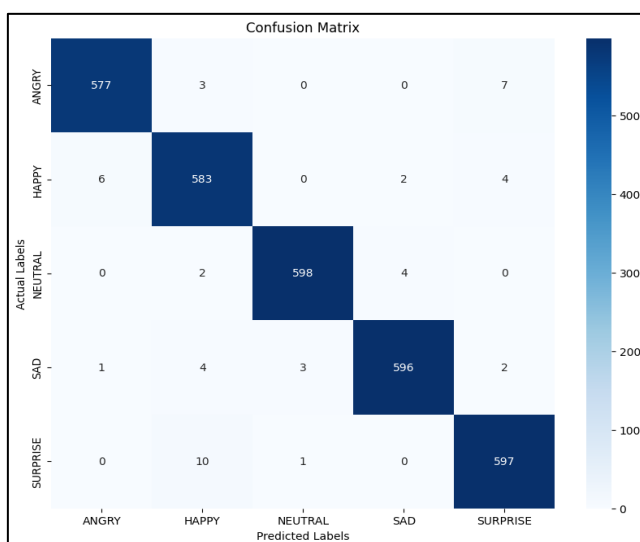
Fig 12: CNN Training Curves (After Augmenting BanglaSER Dataset)



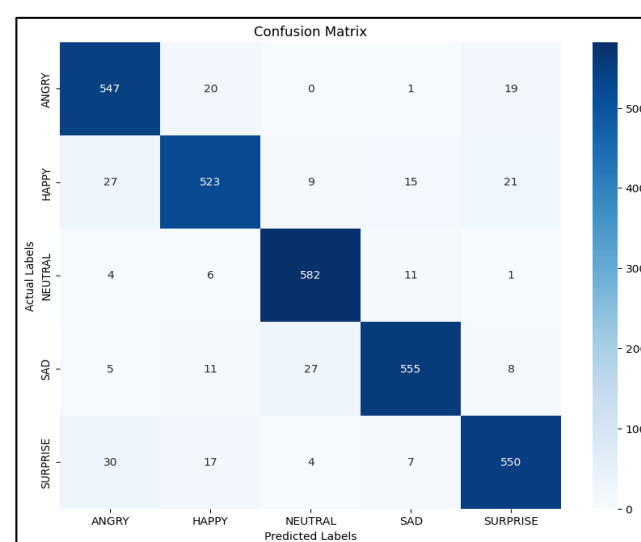
(a) RFC



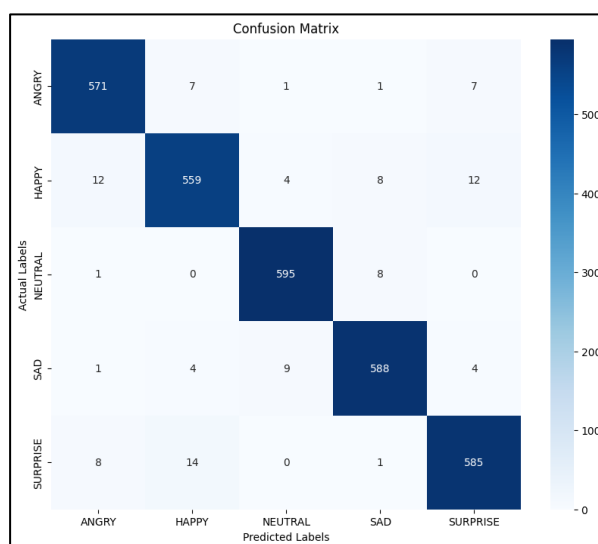
(b) SVM



(c) MLP

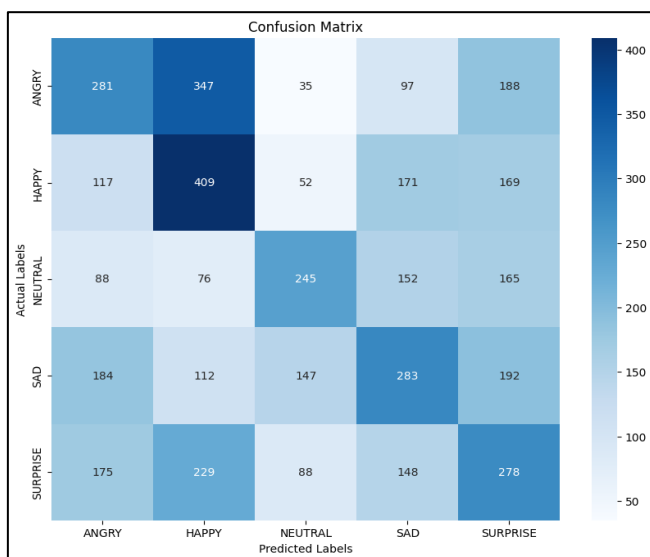


(d) LSTM

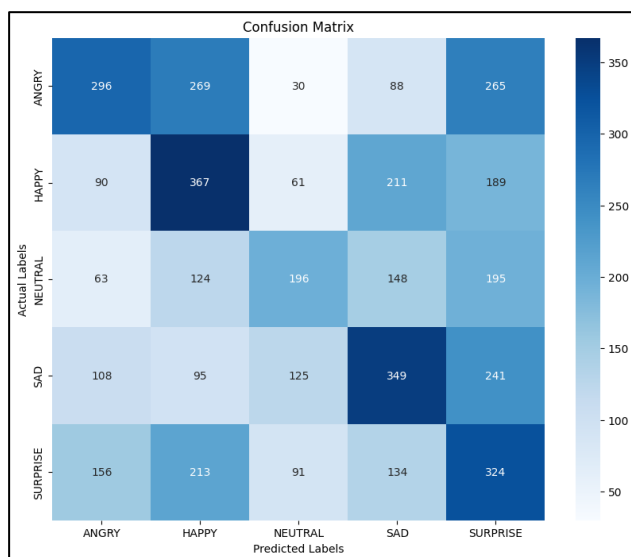


(e) CNN

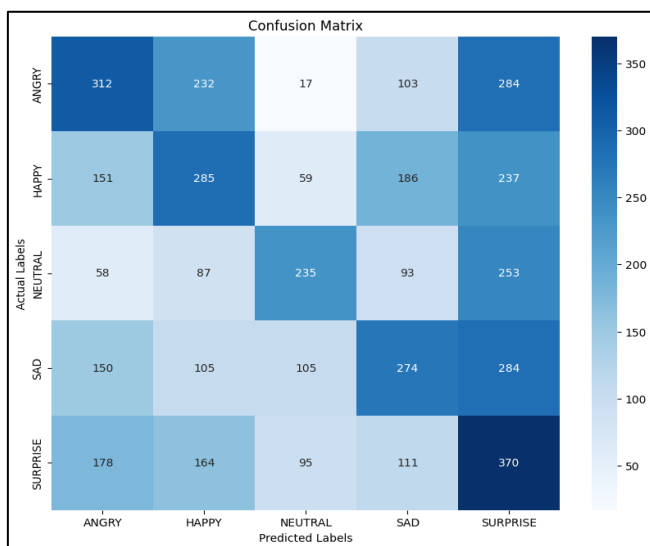
Fig 13: Confusion Matrix of models (SUBESCO dataset)



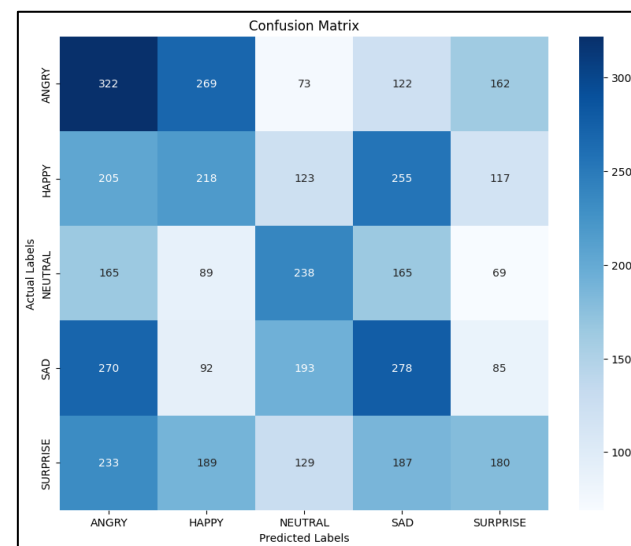
(a) RFC



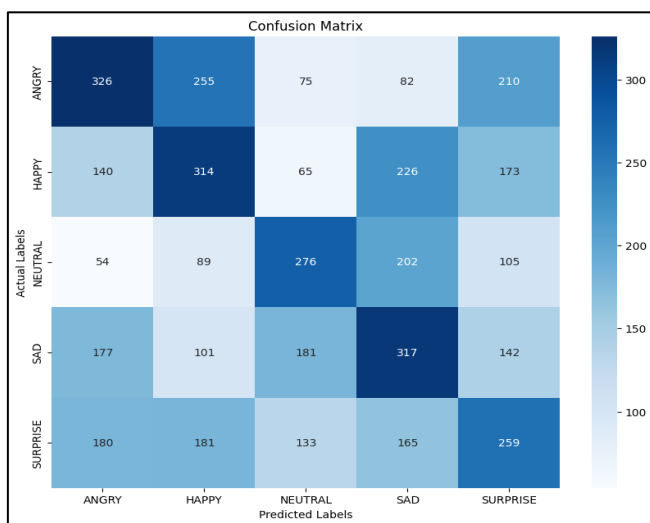
(b) SVM



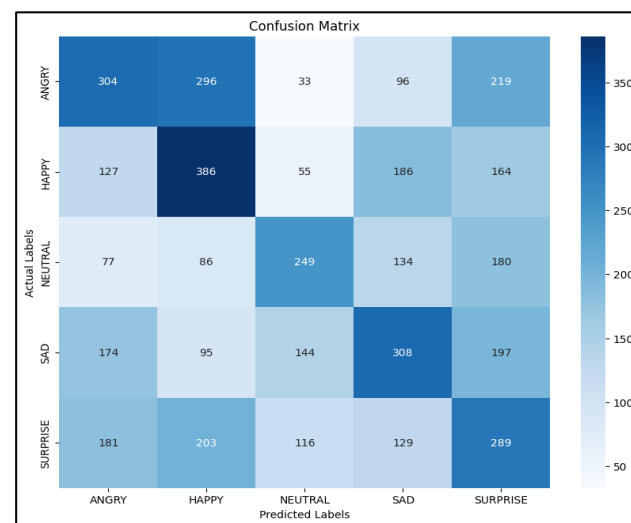
(c) MLP



(d) LSTM

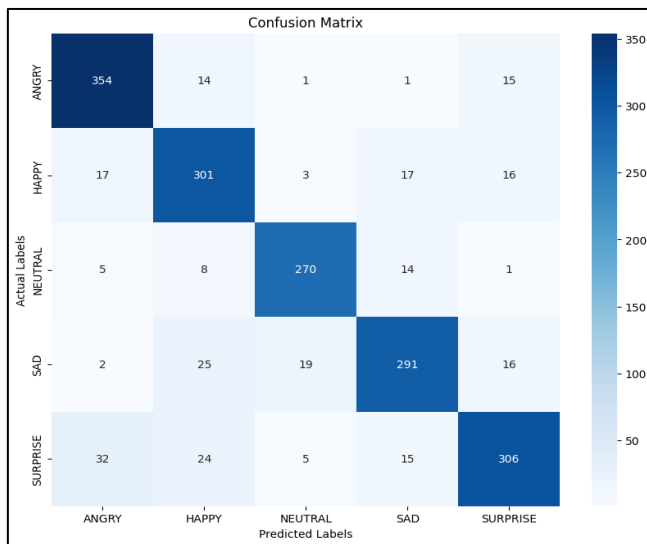


(e) CNN

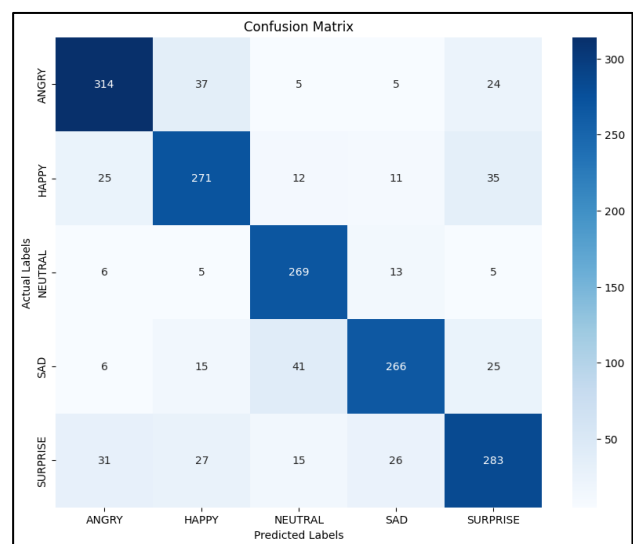


(f) Majority Voting

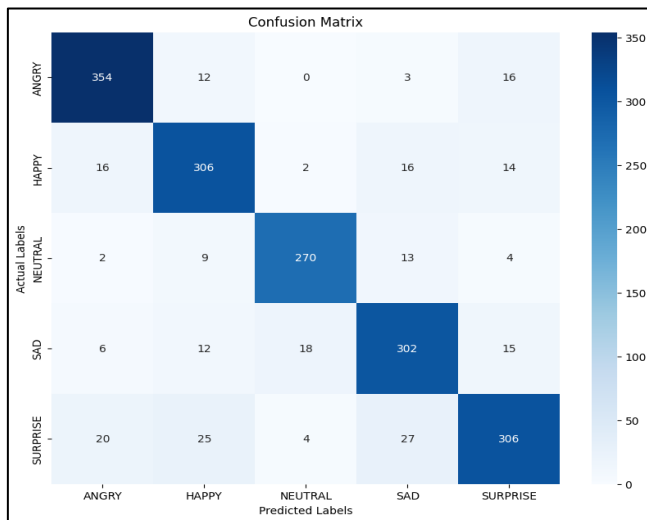
Fig 14: Confusion Matrix of models (BanglaSER dataset without training)



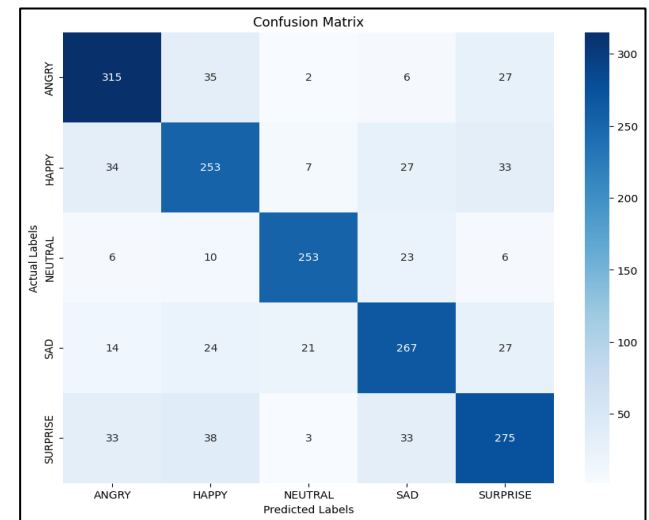
(a) RFC



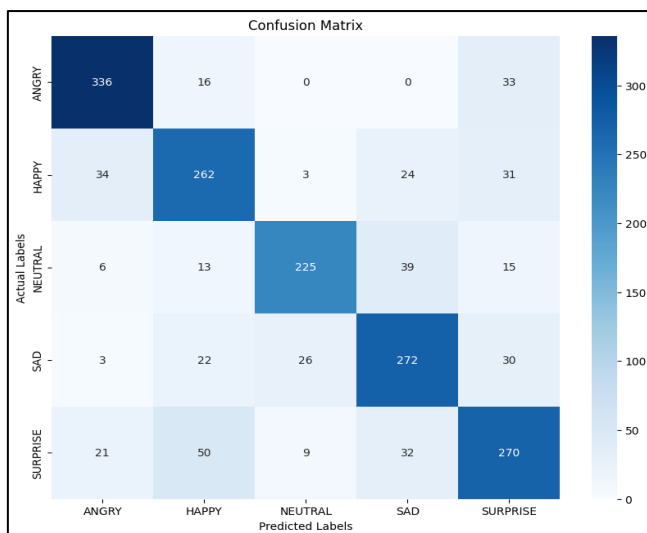
(b) SVM



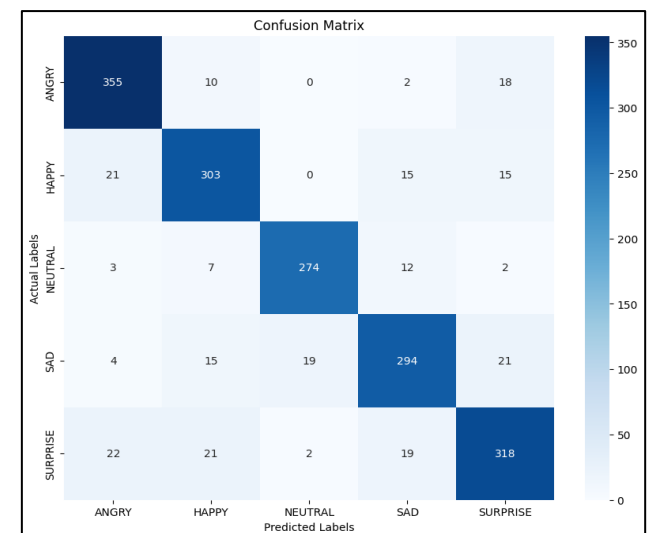
(c) MLP



(d) LSTM



(e) CNN



(f) Majority Voting

Fig 15: Confusion Matrix of models (BanglaSER dataset with training)

As illustrated in Table 4, the Multi-Layer Perceptron (MLP) model provides the best predictions on the SUBESCO dataset. The confusion matrix in Figure 13 reveals that the emotions Happy and Surprise are most frequently misclassified.

However, Table 5 and Figure 14 indicate that our models perform poorly on the BanglaSER dataset without prior training.

Upon retraining with additional data from the BanglaSER dataset, the accuracy nearly doubles, as shown in Table 6. Although the improvement is not dramatic, the MLP model still outperforms the other models. Additionally, the Majority Voting model, which integrates the results of all models, also shows good performance. Nevertheless, the confusion matrix in Figure 16 demonstrates that Happy and Surprise emotions are still miscategorized due to their similar tonal qualities.

2.4. Discussion

In this chapter, we extracted 40 MFCC features and observed that all models achieved an accuracy of approximately 95% on the SUBESCO dataset. However, when evaluated on the BanglaSER dataset without prior training, the models exhibited significantly lower accuracy. This indicates that the models were overfitted to the SUBESCO dataset, lacking generalization capability for new data.

To address this issue, we retrained the models using 60% of the BanglaSER dataset's audio features, out of which 10% features were used for validation of the deep learning models. This retraining resulted in a substantial improvement in performance, with accuracy on the remaining 40% test data rising to around 80% or higher for all models as well as the majority voting model.

This highlights the importance of diverse training data to enhance model robustness and generalization. Future efforts should focus on combining multiple datasets during training to create more comprehensive and adaptable models. Additionally, techniques like cross-validation and data augmentation can further improve the model's ability to generalize to different datasets and real-world scenarios.

CHAPTER 3

STUDY OF BENGALI SPEECH SENTIMENT ANALYSIS USING MEL-FREQUENCY CEPSTRAL COEFFICIENTS FEATURE AND CHROMA SHIFT FEATURE

This chapter provides an overview of the methodologies and results using mfcc and chroma shift feature extraction. Dataset, data preprocessing and data augmentation have been discussed in 2.1, 2.2.2 and 2.2.3 respectively.

3.1. Methodology

3.1.1. Feature Description

MFCC features detail has been discussed in 2.2.1.

Chroma shift [46], is a technique used in audio signal processing to analyze and manipulate the pitch content of a musical signal. This method involves shifting the chroma features, which represent the twelve different pitch classes of the equal-tempered scale in Western music (C, C#/Db, D, D#/Eb, E, F, F#/Gb, G, G#/Ab, A, A#/Bb, B), regardless of octave. Chroma features are essential in music information retrieval tasks as they capture harmonic and melodic characteristics of the audio signal.

The concept of chroma, derived from music theory, has been applied in various digital signal processing applications since the early 2000s. Researchers have utilized chroma features to enhance tasks such as chord recognition, key detection, and cover song identification. By shifting the chroma features, one can transpose the perceived pitch of the audio signal, enabling applications in music genre classification, pitch correction, and audio synthesis.

Chroma shift leverages the robustness of chroma features to alterations in timbre and dynamics, focusing solely on the pitch content. This makes it particularly useful in scenarios where the harmonic structure of the audio is of primary interest. The technique continues to evolve, contributing significantly to advancements in music technology and audio analysis.

3.1.2. Machine Learning and Deep Learning Models

In this project, we utilized traditional machine learning methods including Random Forest Classifier (RFC), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP), alongside deep learning techniques such as Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNNs), for emotion classification. These models are already discussed in detail in 2.2.4 and 2.2.5 respectively.

3.2. Experimental Result

3.2.1. Experimental Setup

The experimental setup and also the steps associated have been elaborated in 2.3.1.

Table 7, 8, 9 presents the accuracy result of the steps respectively.

Figures 16 and 17 display the training curves (accuracy and loss) for LSTM and CNN respectively on SUBESCO dataset.

Figures 18 and 19 display the training curves (accuracy and loss) for LSTM and CNN respectively on BanglaSER dataset.

Figures 20, 21 and 22 illustrate the confusion matrix of the models for the above steps respectively.

3.2.2. Results

Method	Accuracy
RFC	94.17%
SVM	93.07%
MLP	97.50%
LSTM	90.07%
CNN	96.20%

Table 7: SUBESCO dataset (80% training: 20% testing)

Method	Accuracy
RFC	35.21%
SVM	35.61%
MLP	34.60%
LSTM	29.07%
CNN	31.68%
Majority Voting	35.48%

Table 8: BanglaSER dataset (without training)

Method	Accuracy
RFC	84.14%
SVM	78.50%
MLP	84.14%
LSTM	75.06%
CNN	81.32%
Majority Voting	85.72%

Table 9: BanglaSER dataset (60% training: 40% testing)

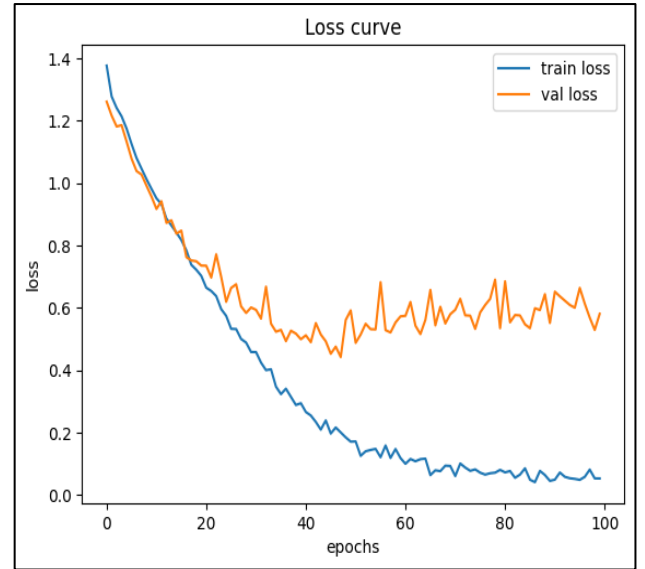
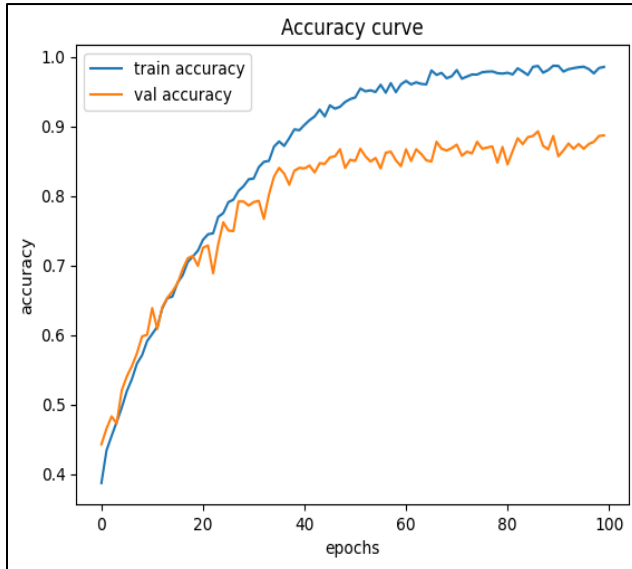


Fig 16: LSTM Training Curves (SUBESCO Dataset)

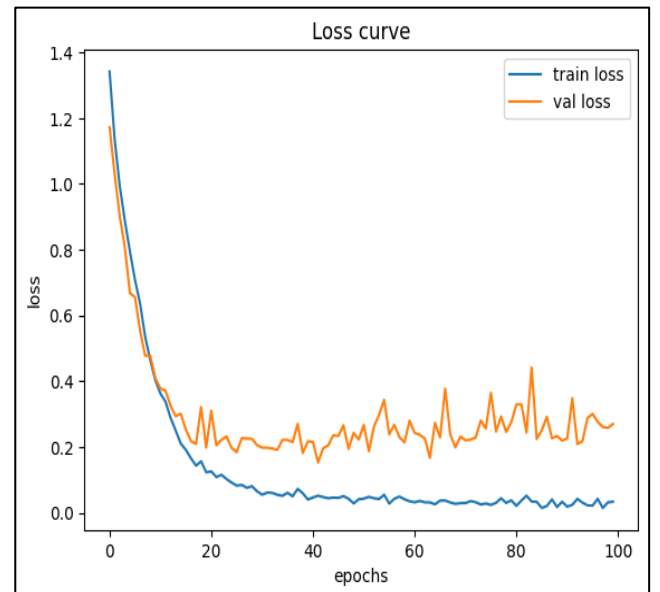
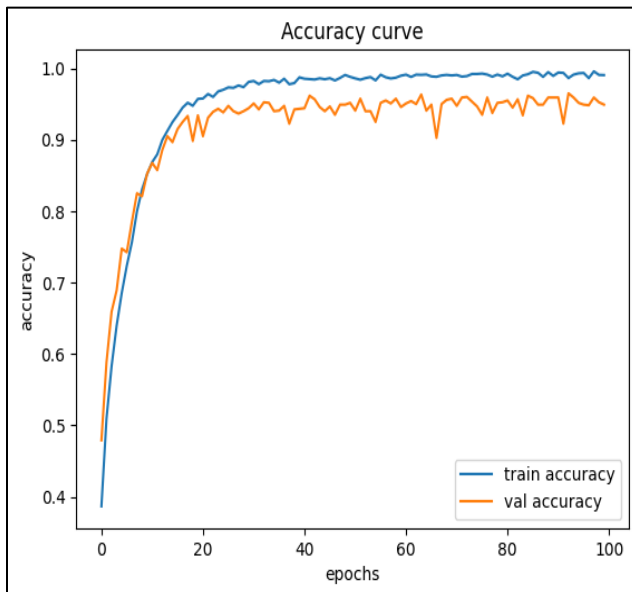


Fig 17: CNN Training Curves (SUBESCO Dataset)

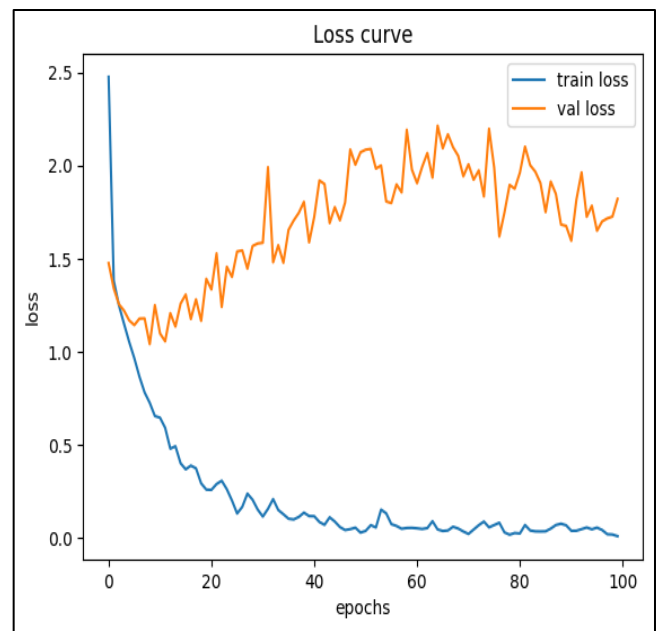
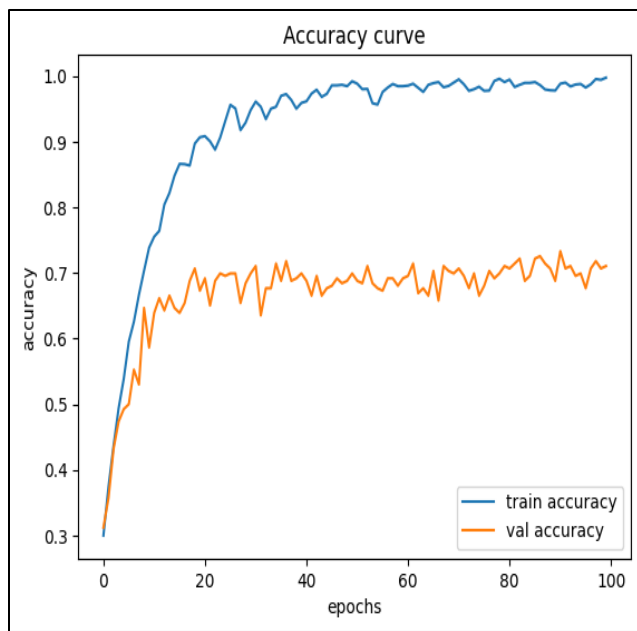


Fig 18: LSTM Training Curves (After Augmenting BanglaSER Dataset)

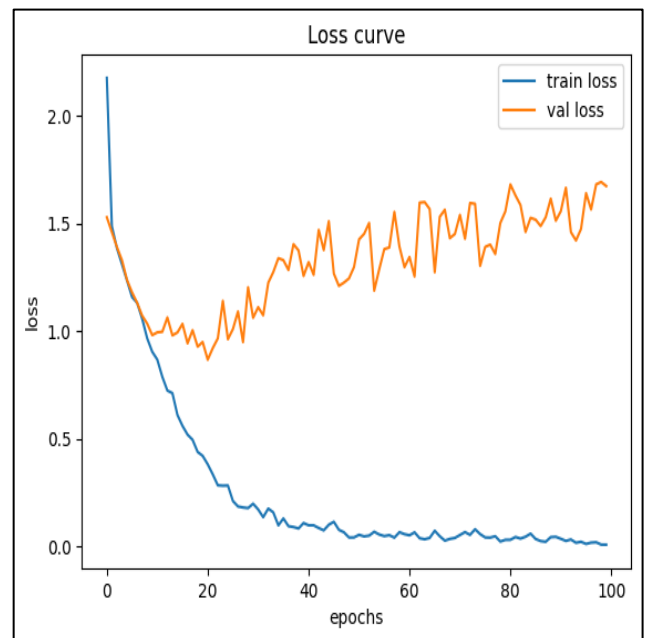
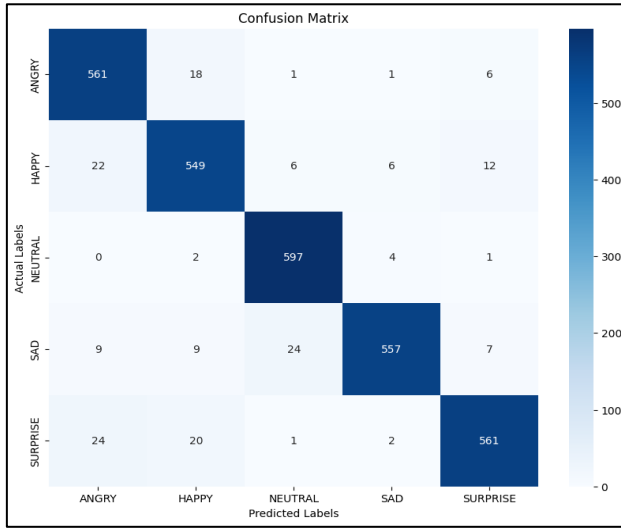
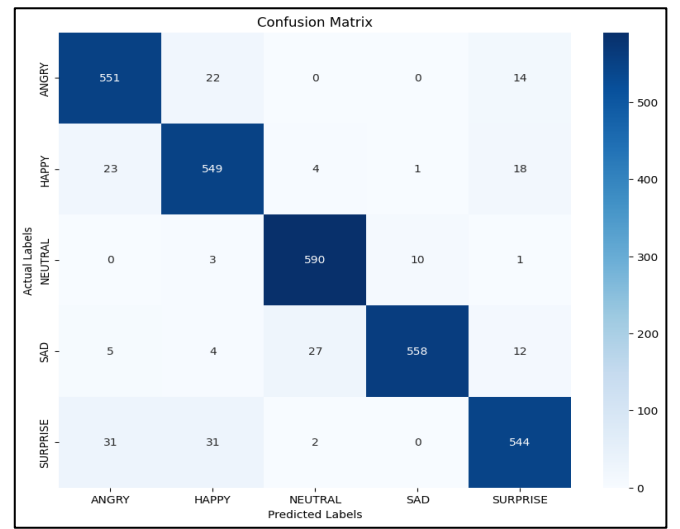


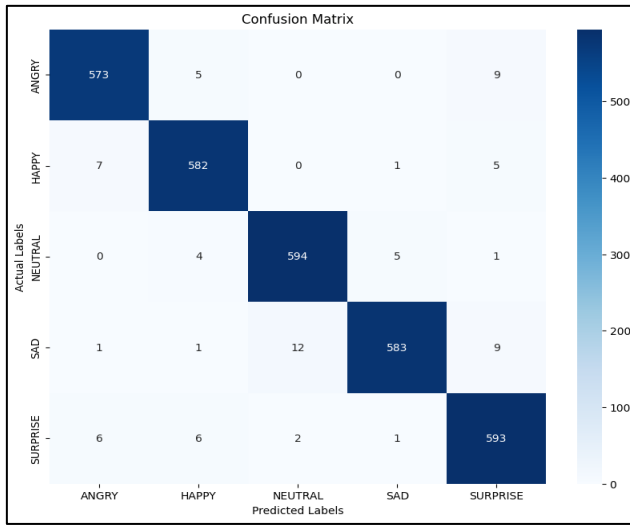
Fig 19: CNN Training Curves (After Augmenting BanglaSER Dataset)



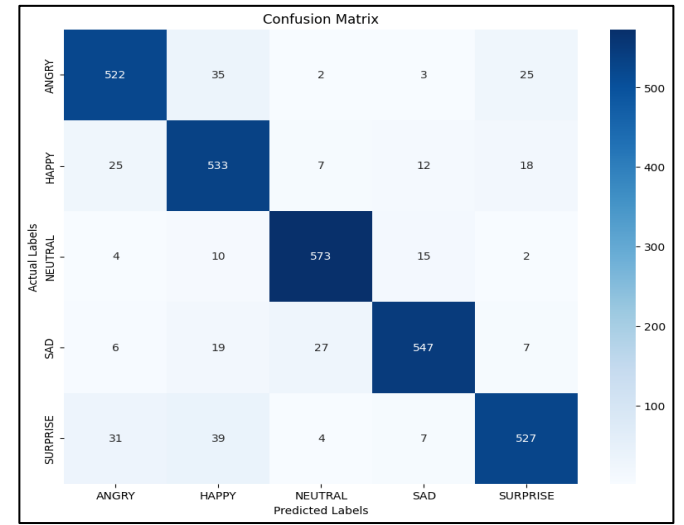
(a) RFC



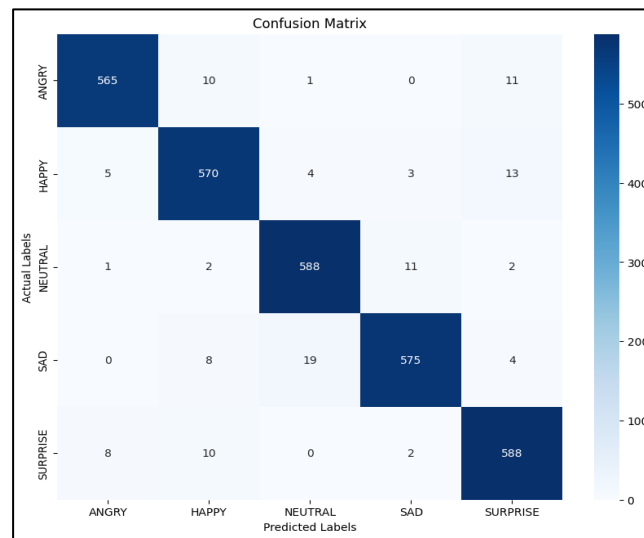
(b) SVM



(c) MLP

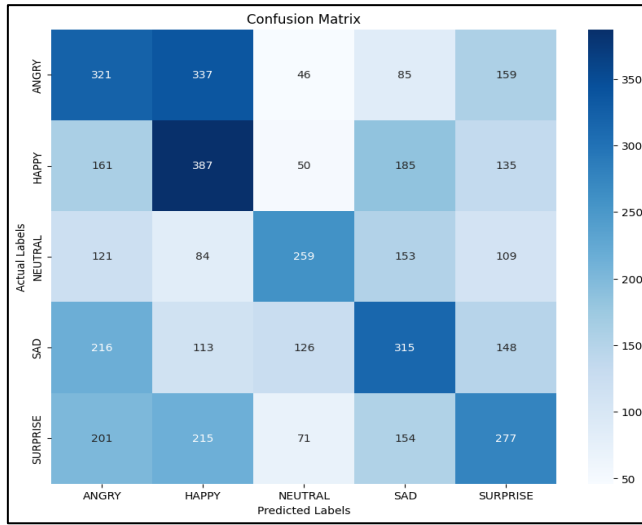


(d) LSTM

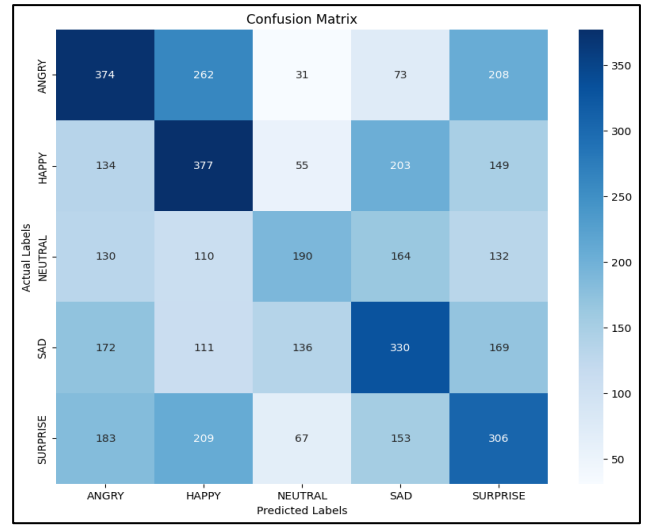


(e) CNN

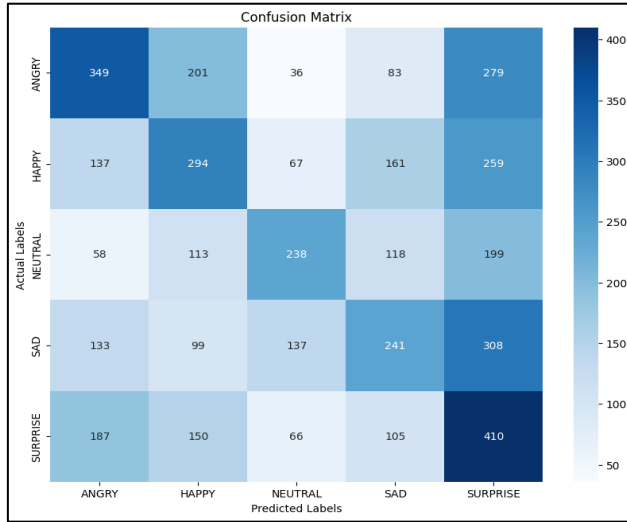
Fig 20: Confusion Matrix of models (SUBESCO dataset)



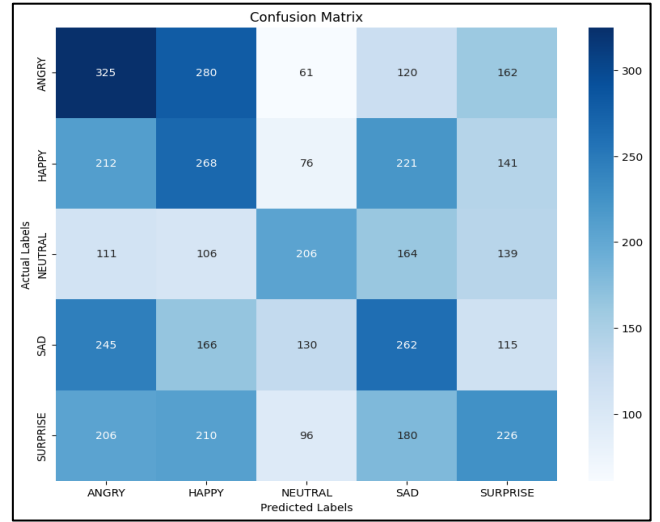
(a) RFC



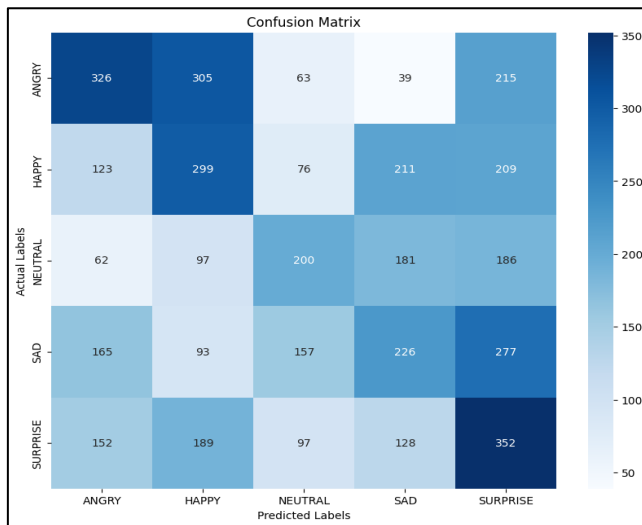
(b) SVM



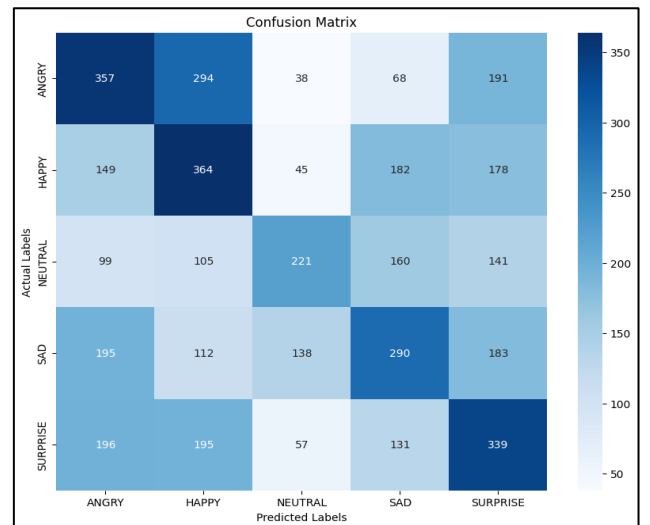
(c) MLP



(d) LSTM

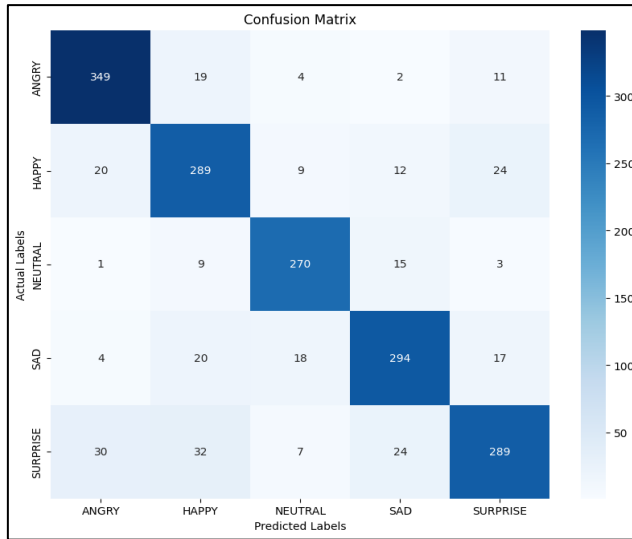


(e) CNN

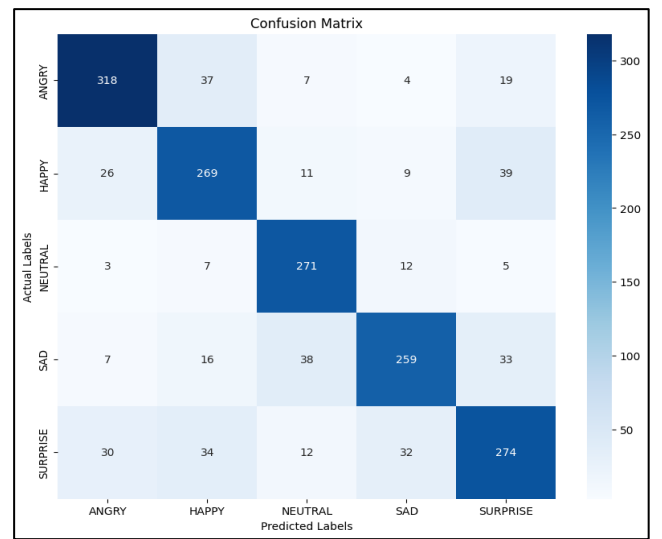


(f) Majority Voting

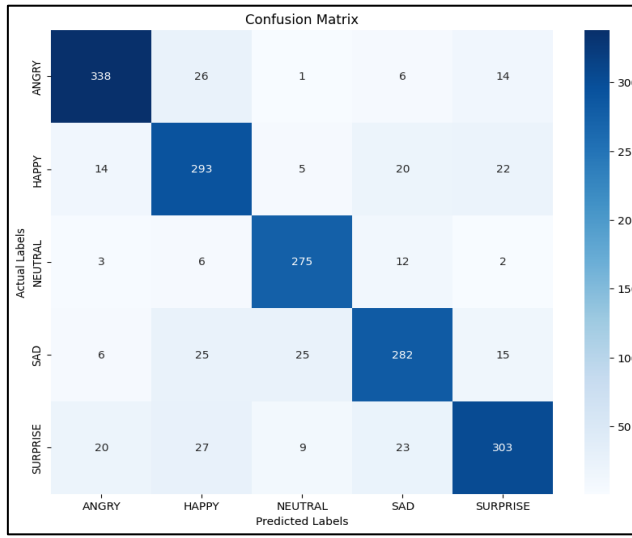
Fig 21: Confusion Matrix of models (BanglaSER dataset without training)



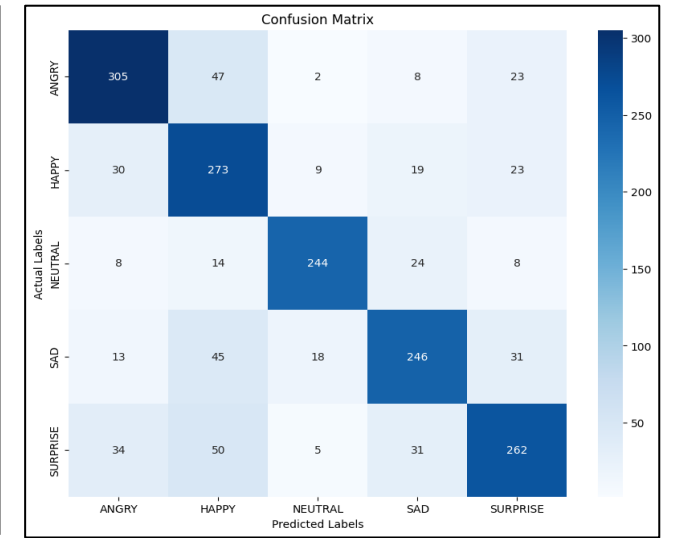
(a) RFC



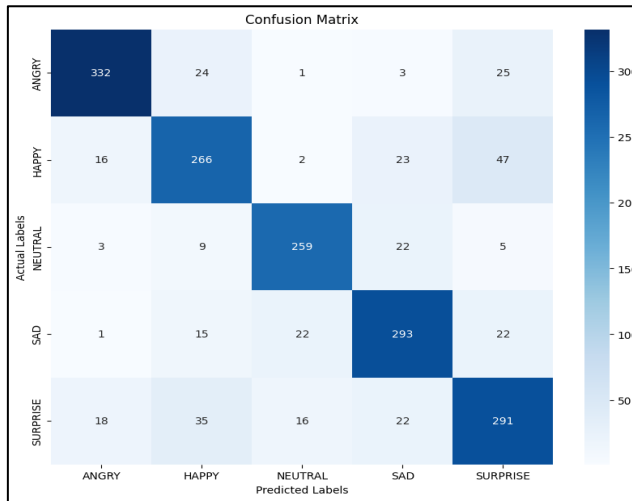
(b) SVM



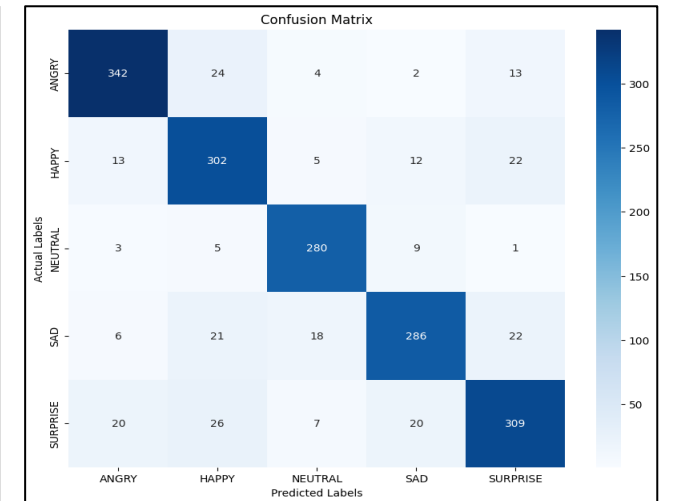
(c) MLP



(d) LSTM



(e) CNN



(f) Majority Voting

Fig 22: Confusion Matrix of models (BanglaSER dataset with training)

As demonstrated in Table 7, the Multi-Layer Perceptron (MLP) model achieves the highest prediction accuracy on the SUBESCO dataset. The confusion matrix in Figure 20 indicates that Happy and Surprise emotions are most commonly misclassified.

In contrast, Table 8 and Figure 21 show that our models perform poorly on the BanglaSER dataset without prior training.

After retraining with additional data from the BanglaSER dataset, the accuracy nearly doubles, as illustrated in Table 9. Despite the moderate improvement, the MLP model still surpasses the other models in performance. Furthermore, the Majority Voting model, which combines the results from all models, also performs well. However, the confusion matrix in Figure 22 reveals that Happy and Surprise emotions continue to be misclassified due to their similar tonal characteristics.

3.3. Discussion

In this chapter, we extracted 40 MFCC features along with 12 additional chroma shift features, which capture tonal differences in scale. We observed that all models achieved an accuracy of approximately 95% on the SUBESCO dataset. However, when evaluated on the BanglaSER dataset without prior training, the models exhibited significantly lower accuracy. This suggests that the models were overfitted to the SUBESCO dataset, lacking the ability to generalize to new data.

To address this issue, we retrained the models using 60% of the BanglaSER dataset's audio features, out of which 10% features were used for validation of the deep learning models. This retraining led to a substantial improvement in performance, with accuracy on the remaining 40% test data increasing to around 80% or higher for all models as well as the majority voting model.

These findings underscore the importance of diverse training data to enhance model robustness and generalization. Future efforts should focus on integrating multiple datasets during training to develop more comprehensive and adaptable models. Additionally, techniques such as cross-validation and data augmentation can further improve the models' ability to generalize to different datasets and real-world scenarios. Employing these strategies will help mitigate overfitting and ensure more reliable performance across varied data environments.

CHAPTER 4

CONCLUSION AND FUTURE WORK

Conclusion

Sentiment analysis based on Bengali speech signals has emerged as a significant topic in human-computer interaction research, driven by its growing application in various fields. In this study, we presented a comparative analysis of sentiment analysis for the Bengali language using both machine learning and deep learning models. We utilized MFCC feature extraction in Chapter 1, and later incorporated MFCC and Chroma Shift feature extraction methods in Chapter 2 to offer a comprehensive comparison between the models.

This project focused on analyzing various machine learning and deep learning approaches for emotion classification in Bengali speech. Through experimental evaluations, we demonstrated the effectiveness of Random Forest, SVM, MLP, LSTM, and CNN algorithms in accurately predicting emotions in Bengali speech samples. The datasets, SUBESCO and BanglaSER, proved to be valuable assets for this task, providing a robust foundation for our analysis.

Our results indicated that the achieved accuracy rates ranged mostly between 85% to 95%, with MLP exhibiting the highest accuracy both with and without additional training. The Majority Voting model also performed well, effectively combining the results from all models to enhance prediction accuracy. This project addresses the scarcity of studies on emotion classification in Bengali, providing valuable insights into the potential for capturing emotions in Bengali speech.

Overall, this emotion prediction framework can assist in understanding individuals' emotional states, enhancing various applications in human-computer interaction and beyond. The insights gained from this study pave way for future explorations and improvements in the domain of Bengali speech sentiment analysis.

Future Work

In this project, our exploration was confined to the SUBESCO and BanglaSER datasets. However, expanding our dataset to include more real-world data sources will be a pivotal step towards augmenting the accuracy and robustness of our models for deployment in real-life scenarios. By integrating diverse datasets, we aim to capture a broader spectrum of variability present in real-world audio recordings, thus enriching the training process and enhancing the generalization capabilities of our models.

While the project predominantly focused on feature extraction techniques such as MFCC and Chroma Shift, the domain of audio signal processing offers many more feature extraction methods, including Mel-spectrogram, root mean square, zero crossing rate, and Fourier transform. An investigation into and incorporation of these additional feature extraction techniques will be instrumental in further refining our models.

In addition to our exploration of traditional machine learning and deep learning models such as RFC, SVM, MLP, LSTM, and CNN, we intend to extend our comparative analysis to include more advanced architectures such as Bi-LSTM and transformer-based models. By evaluating a diverse array of model architectures, we aim to gain deeper insights into their respective strengths and weaknesses, thereby informing our selection of the most suitable model for our specific application domain.

Building upon the foundational findings and methodologies of this project, our future endeavors will focus on the development of a real-time Bengali speech sentiment analysis system. By leveraging the knowledge and insights gained from our research, we aim to create a sophisticated system capable of accurately analyzing and interpreting sentiment from spoken Bengali utterances in real-time. This system will find application in various domains, including customer feedback analysis, market sentiment analysis, and sentiment-aware content recommendation systems, thereby contributing to advancements in natural language processing and human-computer interaction.

REFERENCE

1. B. Mahesh, "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR)*. [Internet], vol. 9, no. 1, pp. 381–386, 2020.
2. M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.
3. R. Verma, V. Nagar, and S. Mahapatra, "Introduction to supervised learning," *Data Analytics in Bioinformatics: A Machine Learning Perspective*, pp. 1–34, 2021.
4. V. Nasteski, "An overview of the supervised machine learning methods," *Horizons. b*, vol. 4, no. 51-62, p. 56, 2017.
5. X. Su, X. Yan, and C.-L. Tsai, "Linear regression," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, no. 3, pp. 275–294, 2012.
6. T. G. Nick and K. M. Campbell, "Logistic regression," *Topics in biostatistics*, pp. 273–301, 2007.
7. D. Boswell, "Introduction to support vector machines," *Departement of Computer Science and Engineering University of California San Diego*, vol. 11, 2002.
8. Y.-Y. Song and L. Ying, "Decision tree methods: Applications for classification and prediction," *Shanghai archives of psychiatry*, vol. 27, no. 2, p. 130, 2015.
9. A. Liaw, M. Wiener, et al., "Classification and regression by randomforest", *R news*, vol. 2, no. 3, pp. 18–22, 2002.
10. Z. Zhang, "Introduction to machine learning: K-nearest neighbors," *Annals of translational medicine*, vol. 4, no. 11, 2016.
11. E. Grossi and M. Buscema, "Introduction to artificial neural networks," *European journal of gastroenterology & hepatology*, vol. 19, no. 12, pp. 1046–1054, 2007.
12. L. Francis, "Unsupervised learning," *Predictive modeling applications in actuarial science*, vol. 1, pp. 280–312, 2014.
13. J. Yadav and M. Sharma, "A review of k-mean algorithm," *Int. J. Eng. Trends Technol*, vol. 4, no. 7, pp. 2972–2976, 2013.
14. M. J. Guess and S. B. Wilson, "Introduction to hierarchical clustering", *Journal of clinical neurophysiology*, vol. 19, no. 2, pp. 144–151, 2002.
15. I.T. Jolliffe, "Principal component analysis: A beginner's guide—i. introduction and application," *Weather*, vol. 45, no. 10, pp. 375–382, 1990.
16. F. Q. Lauzon, "An introduction to deep learning," in *2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, IEEE, 2012, pp. 1438–1439.
17. R. Y. Choi, A. S. Coyner, J. Kalpathy-Cramer, M. F. Chiang, and J. P. Campbell, "Introduction to machine learning, neural networks, and deep learning," *Translational vision science & technology*, vol. 9, no. 2, pp. 14–14, 2020.

18. J. Wu, "Introduction to convolutional neural networks," National Key Lab for Novel Software Technology. Nanjing University. China, vol. 5, no. 23, p. 495, 2017.
19. Á. Zarándy, C. Rekeczky, P. Szolgay, and L. O. Chua, "Overview of cnn research: 25 years history and the current trends," in 2015 IEEE International Symposium on Circuits and Systems (ISCAS), IEEE, 2015, pp. 401–404.
20. R. M. Schmidt, "Recurrent neural networks (rnns): A gentle introduction and overview," arXiv preprint arXiv:1912.05911, 2019.
21. S. Zargar, "Introduction to sequence learning models: Rnn, lstm, gru," Department of Mechanical and Aerospace Engineering, North Carolina State University, 2021.
22. M. Durgadevi et al., "Generative adversarial network (gan): A general review on different variants of gan and applications," in 2021 6th International Conference on Communication and Electronics Systems (ICCES), IEEE, 2021, pp. 1–8.
23. A. Katrekar and B. D. A. AVP, "An introduction to sentiment analysis," GlobalLogic Inc, pp. 1–6, 2005.
24. A. Saxena, H. Reddy, and P. Saxena, "Introduction to sentiment analysis covering basics, tools, evaluation metrics, challenges, and applications," Principles of social networking: the new horizon and emerging challenges, pp. 249–277, 2022.
25. Z. Nanli, Z. Ping, L. Weiguo, and C. Meng, "Sentiment analysis: A literature review," in 2012 International Symposium on Management of Technology (ISMOT), IEEE, 2012, pp. 572–576.
26. B. Nandi, M. Ghanti, and S. Paul, "Text based sentiment analysis," in 2017 International Conference on Inventive Computing and Informatics (ICICI), IEEE, 2017, pp. 9–13.
27. R. Hu, L. Rui, P. Zeng, L. Chen, and X. Fan, "Text sentiment analysis: A review," in 2018 IEEE 4th International Conference on Computer and Communications (ICCC), IEEE, 2018, pp. 2283–2288.
28. J. Breebaart and M. F. McKinney, "Features for audio classification," Algorithms in Ambient Intelligence, pp. 113–129, 2004.
29. F. Pachet and P. Roy, "Exploring billions of audio features," in 2007 international workshop on content-based multimedia indexing, IEEE, 2007, pp. 227–235.
30. I. Luengo, E. Navas, I. Hernáez, and J. Sánchez, "Automatic emotion recognition using prosodic parameters," in Interspeech, Citeseer, 2005, pp. 493–496.
31. J. Devnath, S. Hossain, M. Rahman, H. Saha, M. A. Habib, and N. Sultan, "Emotion recognition from isolated bengali speech," 2020.
32. P. Dhar and S. Guha, "A system to predict emotion from bengali speech," Int. J. Math. Sci. Comput, vol. 7, no. 1, pp. 26–35, 2021.
33. M. R. Islam, A. B. Akhi, F. Akter, et al., "A machine learning approach for emotion classification in bengali speech," International Journal of Advanced Computer Science and Applications, vol. 14, no. 10, 2023.
34. S. Sultana and M. S. Rahman, "Acoustic feature analysis and optimization for bangla speech emotion recognition," Acoustical Science and Technology, vol. 44, no. 3, pp. 157–166, 2023.

35. H. M. Hasan and M. A. Islam, "Emotion recognition from bengali speech using rnn modulation-based categorization," in 2020 third international conference on smart systems and inventive technology (ICSSIT), IEEE, 2020, pp. 1131–1136.
36. M. J. Hasan, M. S. Hossain, S. N. Hassan, M. Al-Amin, M. N. Rahaman, and M. A. Pranjol, "Bengali speech emotion recognition: A hybrid approach using b-lstm," in 2022 4th International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE), IEEE, 2022, pp. 1–7.
37. S. Sultana, M. Z. Iqbal, M. R. Selim, M. M. Rashid, and M. S. Rahman, "Bangla speech emotion recognition and cross-lingual study using deep cnn and blstm networks," IEEE Access, vol. 10, pp. 564–578, 2021.
38. A.C. Shruti, R. H. Rifat, M. Kamal, and M. G. R. Alam, "A comparative study on bengali speech sentiment analysis based on audio data," in 2023 IEEE International Conference on Big Data and Smart Computing (BigComp), IEEE, 2023, pp. 219–226.
39. S. Sultana, M. S. Rahman, M. R. Selim, and M. Z. Iqbal, "Sust bangla emotional speech corpus (subesco): An audio-only emotional speech corpus for bangla," Plos one, vol. 16, no. 4, e0250173, 2021.
40. R. K. Das, N. Islam, M. R. Ahmed, S. Islam, S. Shatabda, and A. M. Islam, "Banglaser: A speech emotion recognition dataset for the bangla language," Data in Brief, vol. 42, p. 108 091, 2022.
41. S. Gupta, J. Jaafar, W. W. Ahmad, and A. Bansal, "Feature extraction using mfcc," Signal & Image Processing: An International Journal, vol. 4, no. 4, pp. 101–108, 2013.
42. P. P. Singh and P. Rani, "An approach to extract feature using mfcc," IOSR Journal of Engineering, vol. 4, no. 8, pp. 21–25, 2014.
43. S. B. S. Meraj and B. Shetty, "Successful data mining: With dimension reduction," in International Conference on Applications of Machine Intelligence and Data Analytics (ICAMIDA 2022), Atlantis Press, 2023, pp. 11–22.
44. A. M. MARCIUC, "Data preprocessing techniques," BUILDING BRIDGES–INNOVATION, ENTREPRENEURSHIP AND SOCIETY, p. 107, 2021.
45. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.
46. M. Kattel, A. Nepal, A. Shah, and D. Shrestha, "Chroma feature extraction," in Conference: chroma feature extraction using fourier transform, vol. 20, 2019.