

Into the Heart of Bangla Speech: Advancing Speech  
Sentiment Recognition with Semi-Supervised Multimodal  
Machine Learning Model Leveraging an Iterative  
SHAP-based Feature Selection

by

Abanti Chakraborty Shruti  
22366034

A thesis submitted to the Department of Computer Science and Engineering  
in partial fulfillment of the requirements for the degree of  
M.Sc. in Computer Science

Department of Computer Science and Engineering  
Brac University  
June 2024

© 2024. Brac University  
All rights reserved.

# Declaration

It is hereby declared that

1. The thesis submitted is my own original work while completing my degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material that has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

**Student's Full Name & Signature:**

---

Abanti Chakraborty Shruti  
22366034

# Approval

The thesis titled “Into the Heart of Bangla Speech: Advancing Speech Sentiment Recognition with Semi-Supervised Multimodal Machine Learning Model Leveraging an Iterative SHAP-based Feature Selection” submitted by

1. Abanti Chakraborty Shruti (22366034)

Of Summer, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of M.Sc. in Computer Science in June, 2024.

## Examining Committee:

Supervisor:  
(Member)

---

Dr. Md. Golam Robiul Alam  
Professor  
Department of Computer Science and Engineering  
BRAC University

Examiner:  
(External)

---

Dr. Shamim H Ripon  
Professor  
Department of Computer Science and Engineering  
East West University

Examiner:  
(Internal)

---

Dr. Md. Ashraful Alam  
Associate Professor  
Department of Computer Science and Engineering  
Brac University

Program Coordinator:  
(Member)

---

Dr. Md. Sadek Ferdous  
Associate Professor  
Department of Computer Science and Engineering  
Brac University

Head of Department:  
(Chair)

---

Dr. Sadia Hamid Kazi  
Associate Professor; Chairperson  
Department of Computer Science and Engineering  
Brac University

# Abstract

Automatic sentiment recognition from speech data is crucial for various applications. As AI has grown in popularity, the application of the importance of speech sentiment analysis is increasing along with the amount of speech in every industry. Bengali is the seventh most spoken language in the world, yet research on voice sentiment analysis in this language is lacking. This thesis investigates novel techniques to enhance speech sentiment recognition in underresourced languages like Bengali. We explore the efficacy of both unimodal (speech only) and multimodal (speech, Image, and text) approaches for different fusion techniques. This research proposed a semi-supervised Random Forest model, which achieved consistent and robust performance across different modality combinations. This model demonstrated high accuracy with fewer features, showcasing the efficiency and effectiveness of SHAP-based semi-supervised learning in handling unlabeled data. Additionally, eight different feature extraction techniques have been employed to extract acoustic features and VGG19 and Bangla Word2Vec are used to extract image and text features. Moreover, this study has experimented with different modality-based methods such as LSTM, CNN, and BanglaBERT. We have used BanglaSER, SUBESCO, and KBES datasets for our experiments. Among the various models tested, early fusion techniques proved the most effective, achieving an accuracy of up to 83% when combining speech and text modalities with LSTM classifiers and the proposed semi-supervised model acquired the highest 77% accuracy for audio, text, and image modals. In contrast, late fusion techniques showed reduced performance, though including speech and image modalities improved accuracy to 62%. Detailed performance comparisons for unimodal systems indicate that traditional Random Forest models perform well with fully labeled datasets, but our semi-supervised model works comparatively well with only 20% labeled data. Moreover, our proposed semi-supervised AdaBoost model, using only 20 features and SHAP-based feature importance, outperformed the traditional model trained with 50 features. Remarkably, the proposed Random Forest model trained with 20% labeled and 80% unlabeled data achieved over 70% accuracy across different feature selection methods, with the weighted feature selection technique achieving the highest accuracy of 72%. We believe this thesis will contribute significantly to Bangla speech sentiment recognition by providing a robust, efficient, and interpretable framework that merges the strengths of deep learning and machine learning models.

**Keywords:** Sentiment Analysis; Machine Learning; Bengali Speech; Emotion; Prediction; Decision tree; AdaBoost; Random Forest; Multimodal; CNN; LSTM; BanglaBERT

## Acknowledgement

First and foremost, I would like to thank the Almighty God for His blessings to keep me healthy and sound throughout the dissertation and for giving me the capability to complete this research work.

Then I would like to express my deep and sincere gratitude to my thesis supervisor, “Dr Md. Golam Robiul Alam” Sir for providing me the opportunity to work with him, for guiding me enormously on every stage, and for being patient with me throughout my research work. Without his immense knowledge, vision, idea formation, resources, and motivation this research work would not have been possible. I would be ever grateful for his teachings and continuous support.

Next, I am very grateful to my family for their unconditional love, care, sacrifices, and prayers for making me able to complete my M.Sc degree and consistently providing me with the mental support to complete this dissertation. Besides, I would like to express my appreciation to my friends for being there whenever I needed them.

I would also like to thank all the faculty members and staff of the Computer Science and Engineering Department for always being thoughtful towards me and supporting me directly and indirectly during my Masters.

# Table of Contents

<b>Declaration</b>	<b>i</b>
<b>Approval</b>	<b>ii</b>
<b>Abstract</b>	<b>iv</b>
<b>Acknowledgment</b>	<b>v</b>
<b>Table of Contents</b>	<b>vi</b>
<b>List of Figures</b>	<b>ix</b>
<b>List of Tables</b>	<b>xi</b>
<b>Nomenclature</b>	<b>xii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Importance of Speech . . . . .	1
1.2 Sentiment Analysis . . . . .	1
1.3 Importance of Speech Sentiment Analysis . . . . .	2
1.4 Bangla Speech Sentiment Analysis . . . . .	2
1.5 Motivation . . . . .	3
1.5.1 The motivation behind using Machine Learning Models for Sentiment Analysis . . . . .	3
1.5.2 Motivation for Feature Selection using SHAP . . . . .	3
1.5.3 Tackling the problem of unlabeled data with Semi-Supervised Learning . . . . .	4
1.5.4 Motivation for Using Multimodal Models . . . . .	4
1.5.5 Motivation for Using Deep Learning Techniques for Multi- modalities . . . . .	5
1.6 Research Gaps . . . . .	5
1.6.1 Lack of Annotated Data . . . . .	5
1.6.2 Exploration of Lightweight Machine Learning Algorithms . . . .	6
1.6.3 Feature Selection with SHAP Technique . . . . .	6
1.6.4 Multimodality Techniques . . . . .	6
1.7 Research Contribution . . . . .	6
1.8 Research Organization . . . . .	7

<b>2</b>	<b>Related Work</b>	<b>8</b>
2.1	Unimodal Speech Sentiment Analysis . . . . .	8
2.1.1	Speech Sentiment Analysis in Bangla . . . . .	8
2.1.2	Bangla Audio Classification . . . . .	9
2.1.3	Speech Sentiment Analysis in Different Languages . . . . .	10
2.2	Multimodal Sentiment Analysis(MSA) . . . . .	12
2.2.1	Bangla Multimodal Systems . . . . .	13
2.2.2	Multimodal Sentiment Analysis Systems in Different Languages	13
<b>3</b>	<b>Methodology</b>	<b>19</b>
3.1	Dataset . . . . .	20
3.1.1	SUBESCO . . . . .	20
3.1.2	BanglaSER . . . . .	21
3.1.3	KBES (KUET Bangla Emotion Speech) . . . . .	23
3.1.4	Data Aggregation . . . . .	23
3.1.5	Dataset Distribution . . . . .	24
3.2	Data Augmentation . . . . .	26
3.3	Dataset Split . . . . .	29
3.4	Feature Extraction for Audio Modality . . . . .	29
3.4.1	Mel frequency cepstral coefficients (MFCC) . . . . .	29
3.4.2	Zero Chroma Rate (ZCR) . . . . .	30
3.4.3	Chroma Shift . . . . .	31
3.4.4	Root Mean Square (RMS) . . . . .	31
3.4.5	Spectral Centroid . . . . .	31
3.4.6	Spectral Bandwidth . . . . .	32
3.4.7	Spectral Roll-off . . . . .	32
3.4.8	Spectral Flatness . . . . .	32
3.5	Feature Generation for Image Modality . . . . .	33
3.5.1	Melspectrogram Generation . . . . .	33
3.5.2	VGG19 Training . . . . .	33
3.6	Feature Generation for Text Modality . . . . .	35
3.6.1	Audio to Text Transcribe . . . . .	35
3.6.2	Generating Embedding Feature with BengaliWord2Vec . . . . .	35
3.6.3	Encoding for Transformer Input . . . . .	36
3.7	Feature Selection . . . . .	36
3.8	Re-sampling and Normalization . . . . .	37
3.9	Multimodal Models . . . . .	37
3.9.1	Early Fusion Model . . . . .	37
3.9.2	Late Fusion Models . . . . .	40
3.9.3	Model Training Details . . . . .	41
3.10	Unimodal Models . . . . .	45
3.10.1	Random Forest Model . . . . .	46
3.10.2	AdaBoost Model . . . . .	46
<b>4</b>	<b>Results &amp; Discussions</b>	<b>47</b>
4.1	Performance Metrics . . . . .	47
4.1.1	Accuracy . . . . .	48
4.1.2	Precision . . . . .	48
4.1.3	Recall . . . . .	48



4.1.4	F-1 Score . . . . .	48
4.2	Experimental Setup . . . . .	49
4.3	Ablation Study . . . . .	49
4.3.1	Feature Selection . . . . .	49
4.3.2	Results of Unimodal Systems . . . . .	49
4.4	Results of Multimodal Systems . . . . .	54
4.4.1	Audio Text Modal . . . . .	54
4.4.2	Audio Image Modal . . . . .	58
4.4.3	Audio Text Image Modal . . . . .	64
4.4.4	Combined Result Analysis on Test Set . . . . .	70
4.4.5	Comparison of Random Forest and LSTM Model for Semi-Supervised Learning . . . . .	72
4.5	Discussion . . . . .	73
<b>5</b>	<b>Conclusion</b>	<b>74</b>
5.1	Limitation and Future Work . . . . .	75
	<b>Bibliography</b>	<b>81</b>
	<b>Appendix A</b>	<b>82</b>

# List of Figures

3.1	Proposed Early Fusion Semi-supervised Multimodal ML Approach with Iterative Feature Boosting using SHAP . . . . .	19
3.2	Proposed Late Fusion Multimodal Approach System Diagram . . . . .	20
3.3	SUBESCO Dataset Data Samples . . . . .	21
3.4	BanglaSER Dataset Data Samples . . . . .	22
3.5	KBES Dataset Data Samples . . . . .	23
3.6	Data Distribution of Final Dataset . . . . .	26
3.7	Generated MelSpectrograms for Positive, Negative and Neutral Classes	34
3.8	Example of Transcribed Text from Audio Files . . . . .	35
3.9	Process of Generating Embedding from Speech . . . . .	36
3.10	Process of Generating Encoding from Speech . . . . .	36
3.11	Proposed Early Fusion LSTM Model . . . . .	38
3.12	LSTM Structure with Equations . . . . .	40
3.13	Custom CNN for generating Visual Features . . . . .	41
3.14	BERT structure (Image taken from [68]) . . . . .	42
3.15	Proposed Late Fusion Model for Audio and Text . . . . .	43
3.16	Proposed Late Fusion Model for Audio and Image . . . . .	44
3.17	Proposed Late Fusion Model for Audio, Text and Image . . . . .	45
4.1	Confusion Matrix for Evaluation Metrics . . . . .	47
4.2	Confusion Matrix for Basic Random Forest Model . . . . .	50
4.3	Confusion Matrix for Basic AdaBoost Model . . . . .	50
4.4	Confusion Matrix for Weighted Random Forest Model . . . . .	51
4.5	Confusion Matrix for Weighted AdaBoost Model . . . . .	51
4.6	Random Forest Training Evaluation Metrics for Audio Text Modal . .	54
4.7	Confusion Matrix of Random Forest for Audio Text Modal . . . . .	55
4.8	Early Fusion Training vs Validation Accuracy for Audio Text Modal .	56
4.9	Early Fusion Training and Validation Loss for Audio Text Modal . .	56
4.10	Confusion Matrix of Early Fusion Technique for Audio Text Modal .	57
4.11	Late Fusion Technique Training vs Validation Accuracy Evaluation Metrics for Audio Text Modal . . . . .	58
4.12	Confusion Matrix of Late Fusion Technique for Audio Text Modal . .	59
4.13	Random Forest Training Evaluation Metrics for Audio Image Modal .	60
4.14	Confusion Matrix of Random Forest for Audio Image Modal . . . . .	60
4.15	Early Fusion Training vs Validation Accuracy for Audio Image Modal	61
4.16	Early Fusion Training and Validation Loss for Audio Image Modal . .	61
4.17	Confusion Matrix of Early Fusion Technique for Audio Image Modal .	62

4.18	Late Fusion Technique Training vs Validation Accuracy Evaluation Metrics for Audio Image Modal . . . . .	63
4.19	Confusion Matrix of Late Fusion Technique for Audio Image Modal .	63
4.20	Random Forest Training Evaluation Metrics for Audio Text and Image Modal . . . . .	64
4.21	Confusion Matrix of Random Forest for Audio Text and Image Modal	65
4.22	Early Fusion Training vs Validation Accuracy for Audio, Text and Image Modal . . . . .	66
4.23	Early Fusion Training and Validation Loss for Audio, Text and Image Modal . . . . .	67
4.24	Confusion Matrix of Early Fusion Technique for Audio, Text and Image Modal . . . . .	68
4.25	Late Fusion Technique Training vs Validation Accuracy Evaluation Metrics for Audio Text and Image Modal . . . . .	68
4.26	Confusion Matrix of Late Fusion Technique for Audio, Text and Image Modal . . . . .	69
4.27	Test Accuracy Comparison Among Different Models . . . . .	72

# List of Tables

2.1	Literature Summary of Bengali Language Data . . . . .	10
2.2	Literature Summary for Unimodal Methods . . . . .	12
2.3	Literature Summary for Multimodal Methods . . . . .	15
2.4	Summary of the State-of-the-arts Speech Sentiment Analysis . . . . .	16
3.1	Dataset Distribution Details . . . . .	25
3.2	Example of Data Augmentation for All Three Datasets . . . . .	28
3.3	Hyperparameter Setting for Vgg19 Model Training . . . . .	35
3.4	Hyperparameters of Early Fusion Model Training . . . . .	42
3.5	Hyperparameters of Audio Text Modal Late Fusion Model Training .	43
3.6	Hyperparameters of Audio Image Modal Late Fusion Model Training	44
3.7	Hyperparameters of Audio, Text, and Image Modal Late Fusion Model Training . . . . .	45
4.1	Results of Feature Selection Test . . . . .	49
4.2	Overall Result Comparison of Unimodal Models on Test Set . . . . .	53
4.3	Classification Report of Proposed ML Model for Multimodal(Audio and Text) Approach . . . . .	55
4.4	Classification Report of Early Fusion Deep Learning Model for Mul- timodal(Audio and Text) Approach . . . . .	57
4.5	Classification Report of Late Fusion Model for Multimodal(Audio and Text) Approach . . . . .	58
4.6	Classification Report of Proposed ML Model for Multimodal(Audio and Image) Approach . . . . .	59
4.7	Classification Report of Early Fusion Deep Learning Model for Mul- timodal(Audio and Image) Approach . . . . .	62
4.8	Classification Report of Late Fusion Model for Multimodal(Audio and Image) Approach . . . . .	64
4.9	Classification Report of Proposed ML Model for Multimodal(Audio, Text and Image) Approach . . . . .	66
4.10	Classification Report of Early Fusion Deep Learning Model for Mul- timodal(Audio, Text and Image) Approach . . . . .	67
4.11	Classification Report of Late Fusion Model for Multimodal(Audio, Text and Image) Approach . . . . .	69
4.12	Combined Result Analysis on Test Set for Multimodal System . . . . .	71
4.13	Performance Comparison between RF and LSTM for Semi-Supervised Multimodal Model (Audio, Image and Text) . . . . .	72
5.1	Grid Search Results for AdaBoost . . . . .	83

5.1	Grid Search Results for AdaBoost . . . . .	84
-----	--	----

# Chapter 1

## Introduction

Sentiment analysis for different language modalities specifically in Bangla has garnered significant interest recently. However, due to limited resources, the experiments on sentiment recognition are complex. This study proposes enhanced machine learning approaches like Random Forest and AdaBoost with iterative feature boosting in a semi-supervised learning loop by leveraging the feature importances extracted from SHAP values to combat unlabelled data and enable feature dimension reduction. Moreover, it develops a comprehensive exploration of unimodal and multimodal methodologies for Bangla Speech sentiment recognition by using state of the art neural architectures like Long-Short time Memory(LSTM) for audio modality, Convolutional Neural network (CNN) for image modality and transformer-based BanglaBERT for text modality with early and late fusion techniques.

### 1.1 Importance of Speech

People's everyday lives depend heavily on communication, where speaking conveys both verbal comprehension and emotional content. The emotions are strongly ingrained regarding voice composition, tone, frequency, pitch, and other aspects of the speech. While it is difficult to have machines comprehend human language in text, it is much more important and necessary to make machines comprehend the emotion and meaning that humans are expressing. In the age of automation, more people are sending voice notes rather than texts for daily correspondence, and they are doing more tasks using voice commands alone. As a result, speech-processing researchers are finding that understanding the emotions that underlie speech is increasingly crucial. Understanding the different attributes is also needed to get more accurate results.

### 1.2 Sentiment Analysis

Sentiment analysis is the study of analyzing and determining an individual's feeling or behavior towards something mostly done within text, audio speech, biometrics, etc. The main motive of sentiment analysis is identifying the expressed opinion's underlying sentiment, whether positive, negative, or neutral. Sentiment analysis is highly important in various sectors such as customer service, product development, brand monitoring, market research, and so on. Understanding and analyzing

the sentiment of customers and stakeholders for a particular product or service can provide valuable insights that can impact future decisions and improve overall performance. Such as sentiment analysis in education [5] is used to improve the teaching quality for students, to develop artificial intelligence [28] to infer human emotions, etc.[55] Moreover, the amount of voice data is vast these days for the same reason. It presents an opportunity for organizations to use the data to improve the quality of their products and develop strategies according to user's intent.

### **1.3 Importance of Speech Sentiment Analysis**

Textual data has been used for the majority of sentiment analysis research, and as large language models and natural language model (NLP) advancements continue, so does the amount of work being done with textual data. Sentiment and emotions do, however, contain certain subliminal indicators that are primarily obscured in the text. Speech data makes it easy to detect these small indicators because certain wordings convey pressure, tone, and pitch. Since voice is the primary mode of communication for most communications, regardless of language or media, speech processing offers the ability to analyze speakers' intent in real-time and has access to a large number of data sources. Therefore, a thorough knowledge of sentiments can be attained if the speech is additionally examined in addition to other modalities like textual data.

### **1.4 Bangla Speech Sentiment Analysis**

While understanding the importance of speech sentiment analysis, most of the researchers have worked on audio data to understand the sentiment features. However, most of the works are done in the English language. As the world is evolving and language is a universal component, the need for machines to learn other low-resource languages such as Bangla is increasing. As of 2021, there were about 240 million native speakers of Bangla and an additional 41 million speakers of the language as a second language. According to [66], Bangla is the sixth most spoken native language globally and the seventh most spoken language overall. It is one of the most commonly spoken languages in the world.[65][18] But Bangla language resources are still inadequate which is why it gets tagged as low resource language. The lack of proper resources makes language-specific tasks like sentiment analysis more complex and this creates a lack of work done for the same. Although few works have been done on Bangla Speech sentiment analysis, the importance of more experimental work in this sector has increased. It is due to the rise of using the Bengali language in social media, political issues, expressing mental health, businesses, e-learning platforms, etc. Therefore, Bengali speech sentiment analysis develops systems that are more intelligent, inclusive, and responsive to the varied requirements of populations speaking Bangla.

## 1.5 Motivation

### 1.5.1 The motivation behind using Machine Learning Models for Sentiment Analysis

For sentiment analysis tasks, mostly the deep learning approaches such as Recurrent Neural Networks, LSTM, Bi-LSTM, attention mechanism, BERT, GRU, and CNN mechanisms, etc are popular in recent times. However, novel machine-learning classification algorithms are also popular in this sector. These algorithms are commonly used for classification tasks. Sentiment analysis being a specific classification task achieves great results with these algorithms. Over time algorithms like Random Forest, K- Nearest Neighbour, AdaBoost, XGBoost, LightGBM, etc are used for the same purpose. Random Forest and KNN achieve the highest performance in this regard. They are mostly used due to their effectiveness in handling nonlinear data, less complex architecture, easier handling of imbalance data, etc. Especially Random Forest and AdaBoost are ensemble models that learn from weak classifiers. Moreover, researchers have experimented with them by stacking them and achieved better performances. However, due to the inadequacy of datasets in the Bangla language, the experiments done for Bangla sentiment classification with machine learning models are also inadequate. In the study [3] it is shown that models like Random Forest and Adaboost work best with real-life audio data. Although the performances of the models could be higher due to external noise and other dependables in real-life audios, the novel methods worked better than the neural network models for Bangla Audio.

### 1.5.2 Motivation for Feature Selection using SHAP

A crucial stage in the machine learning process is feature selection, which entails picking the most pertinent and instructive characteristics to minimize overfitting and improve model generalization. The machine learning techniques used in speech sentiment recognition models are taught using features that are taken from audio files. Several characteristics of an audio file may be essential for determining the sentiment conveyed in a speech. Mel frequency cepstrum coefficient (MFCC), root mean square, zero chroma shift, zero crossing rate, spectral centroid, spectral roll-off, spectral flatness, pitch, average frequency, and so on are a few of these characteristics. Additionally, to minimize the dimensionality of the training set, the majority of these extracted features from audio data are chosen using multiple feature selection algorithms. However, in most cases, they are not caused by the training model itself, which might help pinpoint the features that are having an influence. The models produce and respond to the importance of each attribute as they are being trained. Nevertheless, the model retains the global significance of the features that are utilized for overall prediction after the training process is complete. The model will learn more about the crucial features and functions better if we can apply these features' importance in addition to the local significance of each iteration of the model.

Each characteristic is assigned an important value for a particular prediction by the SHAP (SHapley Additive exPlanations) unified framework for prediction interpreta-



tion.[13]. It enhances a machine-learning model’s interpretability and transparency. Understanding how different features impact predictions and how they contribute to overall predictions is made easier by using SHAP. By adopting the SHAP viewpoint, we may include this explicable technique in machine learning models for feature selection and adjust the model’s prediction.

### **1.5.3 Tackling the problem of unlabeled data with Semi-Supervised Learning**

Speech recordings or samples that have not been manually classified with sentiment labels are referred to as ”unlabeled data” in the context of speech recognition. Labeled sentiment data is scarce for research domains, primarily in the field of Bangla research. The complex structure of the Bangla language makes data tagging challenging. Additionally, speech data is subjective by nature and can differ based on the opinions and prejudices of individual annotators. Additionally, the availability of qualified annotators is restricted by domain and language specialization, which affects the performance of sentiment recognition models that are currently in use. Because the models are trained on a small amount of data, evaluating their performance also presents difficulties. An algorithm called semi-supervised learning picks the most instructive examples for labeling with an emphasis on prediction certainty. This approach can be used to prioritize features by including their importance in the model and learning from the labeled data. The unlabeled data can then be labeled based on these characteristics. Because semi-supervised learning mixes a big amount of unlabeled data with a small amount of labeled data to increase model performance, it can be especially useful in situations when there is a shortage of labeled data.

### **1.5.4 Motivation for Using Multimodal Models**

As we hear, see, talk, and read, we are surrounded by a variety of modalities that help us define our view of the world and the situations in which we find ourselves. An ML (machine learning) model that can analyze data from several modalities, such as text, videos, and images, is called a multimodal model. [70] By utilizing the influence of various modalities to produce a human-like knowledge of the behavior of the feature or about a specific scenario, this model seeks to improve performance. Acquiring knowledge from multimodal sources allows recording correspondences between modalities and developing a comprehensive comprehension of natural processes. [14]. Audio-visual speech recognition (AVSR) is among the first applications of multimodal research [1]. Contextual cues and subtle emotions can be misinterpreted due to the ambiguity and lack of context in speech data. The model can compensate for its shortcomings by obtaining a more comprehensive representation of the sentiment conveyed in the speech through the use of other speech-related modalities, such as text and visual representation. The fact that speech data can contain varying linguistic elements based on the background of the speaker and be exceedingly noisy due to background noise is another crucial factor. All of these could impede the ability to comprehend the speech data’s true content clearly and lead to the model performing poorly on the data that hasn’t been viewed. Enhancing hidden features and strengthening system resilience can be achieved by integrating

data from many modalities of the same content.

Coming to fusion techniques, the fusion techniques for multimodal systems can be largely divided into two broad categories - (1) Early fusion and (2) Late fusion. Moreover, these days intermediate fusion and hybrid fusion techniques have taken a front seat in research too. The definitions for the two main fusion techniques are as follows as described in [71]

- Early fusion: In this approach, raw data from different modalities is combined at the input level before being fed into a model. For example, combining text and image data into a single input vector.
- Late fusion: In this approach, data from each modality is processed independently through separate models, and the outputs from these models are then combined at a later stage.

Studies have demonstrated that modalities can work in conjunction and that appropriate data fusion can result in appreciable gains in model performance [14][22]. The effectiveness and caliber of data fusion strategies to be implemented by researchers are largely dependent on selecting, or even designing, the appropriate data fusion technique. Therefore experimenting with different fusion techniques for the multimodal system is important to get the actual performance of the system.

### **1.5.5 Motivation for Using Deep Learning Techniques for Multimodalities**

The integrated characteristics and optimal structures for various modalities—such as audio, text, and image—vary. Different deep learning algorithms have attained state-of-the-art performance for various modalities, while novel machine learning techniques like KNN, Adaboost, Random Forest, etc. perform well with this data. These neural network architectures utilize the data’s intricate properties to identify and extract their complex pattern. Neural networks in multimodal models can be directly fed raw input data, allowing them to discover hidden characteristics in the data and enable more thorough analysis. Moreover, Pre-trained models, such as BERT for text, CNNs for images, and RNNs for sequential data, capture generic features and semantics from large corpora, which can be leveraged for multimodal tasks with limited labeled data, thus reducing the need for extensive data annotation.

## **1.6 Research Gaps**

Bengali speech sentiment analysis is an emerging field in the literature but several research gaps hinder its process. Addressing these gaps can enable the prompt use of the research in real-life applications. The main identified research gaps are discussed below -

### **1.6.1 Lack of Annotated Data**

The lack of annotated data is one of the primary challenges. There is a shortage of large, annotated datasets for Bengali speech, which are essential for training

robust machine learning models. Existing datasets are often small, not diverse, and lack emotional labeling. Very little work was done to handle the annotated data scarcity in the Bengali language. One notable work among them is the deep learning-based audio-text encoding approach proposed by authors in [24]. However, while the traditional lightweight machine learning approaches have shown promising performance with speech data, no significant work has been done to handle data scarcity with these algorithms. This indicates the importance of work done in this sector to handle the annotated data limitation.

### **1.6.2 Exploration of Lightweight Machine Learning Algorithms**

Traditional lightweight machine learning approaches, known for their efficiency and effectiveness with smaller datasets, have not been extensively experimented with in the context of Bengali speech sentiment analysis. These methods could offer significant benefits in terms of computational efficiency and resource management, especially in scenarios where high computational resources are not available. Research into how these approaches can be adapted and optimized for Bengali speech data could provide valuable insights into the field.

### **1.6.3 Feature Selection with SHAP Technique**

The existing literature does not utilize explainability methods as a feature selection approach. SHAP is one of the most used Explainable techniques which can be used in feature selection too. Explainable methods play a crucial role in identifying the most influencing factors/features for the outcome of a model. Thus, incorporating these techniques can effectively help in the feature selection of the model.

### **1.6.4 Multimodality Techniques**

Multimodal analysis, which involves integrating multiple data sources such as audio, text, and visuals, remains an underexplored field in Bengali speech sentiment analysis research. Multimodal approaches can provide a more comprehensive understanding of sentiment by combining different types of information. However, there is a lack of research on these modalities for Bengali speech. Exploring multimodal methods could significantly improve the accuracy and robustness of sentiment analysis models.

Therefore, to bridge the above-mentioned research gaps, our research proposes a novel semi-supervised multimodal approach to leverage speech recognition and handle the data scarcity problem. It also proposes a way of incorporating the SHAP technique to reduce the feature dimension of the model.

## **1.7 Research Contribution**

In this study, we design and propose a novel multimodal semi-supervised machine learning approach with an early fusion technique and leverage XAI as an iterative feature selection approach to recognize Bengali speech sentiment. In particular, we

propose a robust and efficient semi-supervised technique that works well with fewer features and tackles unlabeled data. Moreover, We aim to provide an extensive study on the sentiment recognition models for Bengali Speech by implementing different machine learning and deep learning models with multimodality. The detailed contributions are summarized as follows -

- We introduce a multimodal Bangla speech sentiment recognition method considering three modalities- Audio, Text, and Image. We propose a semi-supervised model using the best-performing ML models from Literature (Random Forest and AdaBoost) to recognize Bengali speech sentiment. Using this semi-supervised method, we tackle the issue of unlabeled data problems in the existing literature with this approach where the model is trained with only 20% of labeled data and learns from the 80% of unlabeled data. Also, Speech-Text, Speech-Image, and Speech-Text-Image multimodal models with two different fusion techniques - Early fusion and Late fusion are experimented with in this research to leverage the modality-based scenario impact of the model.
- We implement an early-fusion method for the multimodal method with an iterative weighted feature-boosting approach integrating the SHAP values into the semi-supervised learning loop to train the model with the most informative samples. This approach identifies only the highest impacting features reducing the dimensionality of the model and reducing the time and resource cost of the model training.
- We also implement a late fusion technique for the multimodal method using the Sequence model(LSTM), transformer-based model(BanglaBERT), and CNN for audio, text, and image modality respectively. Using the late fusion multimodal approach, we aim to make the model learn integrated information from different modalities independently and identify the sentiments effectively.
- We presented a comprehensive performance study of different implemented Unimodal and Multimodal models with a merged SUBESCO, BanglaSER, and KBES dataset regarding evaluation metrics such as Accuracy, Precision, Recall, and F-1 Score. Moreover, we investigated the importance of feature numbers and the impact of feature reduction in speech recognition. Experimental results depict that the proposed multimodal approach performs significantly well with only 20 features and 80% unlabeled data ensuring the model's effectiveness.

## 1.8 Research Organization

This report's remaining content is formatted as follows: In Chapter 2, the background research for this research is discussed. Methodologies are discussed and covered in detail in Chapter 3. Chapter 4 explores the results and analysis of the findings. Chapter 5 presents the main conclusion of the thesis.

# Chapter 2

## Related Work

The literature review on speech sentiment analysis spans methodologies applied to various languages, including Bengali, and encompasses both unimodal and multimodal approaches. It explores the use of machine learning and deep learning techniques to discern emotional cues conveyed through spoken language. The review investigates challenges such as linguistic nuances, dataset availability, and model interpretability, while also examining the integration of visual modalities for improved sentiment analysis outcomes. By exploring advancements and limitations in speech sentiment analysis across languages, this review aims to identify trends and gaps and highlight its contribution and possible future research directions in the field.

### 2.1 Unimodal Speech Sentiment Analysis

Unimodal sentiment analysis stands as one of the most prolific areas of research, witnessing extensive exploration across multiple languages. Numerous approaches, from conventional machine learning algorithms to state-of-the-art deep learning architectures, have been adopted in this dynamic sector. Notably, recent advancements in the field have been propelled by the emergence of transformer and transfer learning models, revolutionizing the landscape of sentiment analysis across diverse linguistic contexts.

#### 2.1.1 Speech Sentiment Analysis in Bangla

Most of the sentiment analysis tasks of the Bangla language were focused on texts while few works have examined the Speech Sentiment Analysis for the Bangla language; one notable example is the study in [31], which suggested the DCTFB architecture used for Speech Emotion Recognition, which combines a TDF layer with Deep CNN and Bi-LSTM. The RAVDESS and SUBESCO which is the audio-only Bangla sentiment speech dataset are used in the trials. When it comes to emotional speech, the model can learn both local and sequential information. After testing eight models, the DCTFB model turned out to be the best. Using SUBESCO dataset it achieved an accuracy(weighted average) of 86.86% and a f1 score(average) of 86.86%. The model's accuracy with the RAVDESS dataset was 82.7% WA. A different study [27] suggested a method for identifying emotions in Bengali speech. To extract features from the speech stream, MFCC and LPC are coupled. This system uses several machine learning methods, including SVM, KNN, AdaBoost,

Logistic Regression, and XGBoost, to predict sentiment; LR and SVM performed the best. The system’s performance is evaluated using two datasets: the well-known RAVDESS dataset and Abeg a dataset of self-compiled audio recordings that contains 301 audio speeches. On the Abeg dataset, the Logistic Regression model had the highest accuracy (92%). Using the combined two datasets - RAVDESS and Abeg datasets, the XGBoost classifier produced an 86% accuracy rate for twelve users.

In the study [15], an automatic speech recognition system with an accuracy of 86.08% is suggested to identify isolated Bengali words. This model uses SVM with dynamic time wrapping (DTW). While DTW is used for feature matching, MFCC is utilized to detect the features of audio. Afterwards, this model uses SVM for classification based on these features. The study employed a self-collected dataset of 40 distinct speakers’ pronunciations of five Bangla terms for training. In a study [55] authors made a comparison between different ML and DL techniques to show the performance difference between models and explained the results with explainability. Recent works such as [48] authors focused on the comparison of audio features and models for multitask Bangla audio analysis by using Subesco and Ravdess datasets with MLP, Random Forest, Gradient Boosting, and SVM models their findings generated that MFCC and Chroma features are highly effective for gender speaker and emotion recognition. A custom 1d CNN-based approach is proposed by authors in [45] where 90% accuracy is achieved for the SUBESCO dataset by utilizing MFCC features. A Bi-LSTM model was employed using MFCC, Chroma, and Mel-Spectrogram features in [38] where the experiments achieved an accuracy of 83.33% for seven class classifications.

### 2.1.2 Bangla Audio Classification

Jahid et al. [29] propose a fresh dataset of 5120 audio clips and classification models for the same Bangla audio for the purpose of classifying Bangla audio news. For the same goal, the authors experimented with deep learning, transfer learning, and classic machine learning models. Using transformer architecture and MFCC characteristics, they were able to reach the maximum accuracy of 93.2%. Another approach to recognizing Bangla Short Speech Commands using the CNN model has been taken by the authors Sumon, Shakil et al. [16]. The authors used MFCC feature extraction techniques on their custom dataset containing 65,000 samples and fed the features to raw 1d CNN models and a pre-trained model of Google where they achieved the highest accuracy of 74%.

Recently, there has been an increase in the popularity of music classifications, especially in Bangla, and academics have used several methods. Among them, authors in [39] offered an additional cutting-edge method called BMNet-5 to categorize Bengali music into six categories using integrated audio features. With an accuracy of 90.32%, their proposed model surpassed the matching prior study. They made use of the 1742 music audio files’ MFCC, ZCR, and spectral properties. Similarly, the paper [17] proposes to classify Bengali music genres using a neural network technique. Jibon[53] has conducted more work in this area in which pre-trained transformer

Table 2.1: Literature Summary of Bengali Language Data

Paper	Dataset	Method	Performance(%)
[31]	SUBESCO, RAVDESS	DCTFB	86.86
[27]	RAVDESS, ABEG	SVM,KNN,AdaBoost, Logistic Regression, and XGBoost	92
[15]	Self-collected audio	SVM with dynamic time wrapping (DTW)	86.08
[48]	SUBESCO, RAVDESS	MLP, Random Forest, Gradient Boosting, and SVM	87.54
[45]	SUBESCO, BanglaSER	Custom 1d CNN	90
[38]	SUBESCO	Bi-LSTM	83.33
[29]	BAND	ML, DL, and Transfer learning models	93.2
[39]	Custom Dataset	BM-Net-5	90.32
[17]	Custom Dataset	Deep learning model	70
[53]	BanglaBeats	DistilHubert, Wav2Vec2-Base-960h	84.94
[54]	Dataset of [17]	WaveNet	-
[16]	Custom Dataset	Raw and pretrained models	74

models, such as DistilHubert and Wav2Vec2-Base-960h, were used to identify the genres, with a maximum classification accuracy of 84.94%. Additionally, Khan et al. combined K-Fold with Principal Component Analysis on the WaveNet model in [54]. The summary of all the literature that have used Bengali language data is provided in Table 2.1

### 2.1.3 Speech Sentiment Analysis in Different Languages

To improve the models' performance, deep learning-based models are used in the majority of investigations. The study [44] suggests a deep learning model to improve the prediction rate and accuracy of the current algorithms. This study's primary goal was to enhance the information extraction process for speech features. This work presents a new framework that uses an attention-based GRU model to fuse the spatiotemporal aspects of speech with a hierarchical conformal model to extract those features. Reducing the deep learning model's computational expense for feature extraction was another goal of the model. The performance of the new system is assessed using the IEMOCAP and RAVDESS benchmark datasets, demonstrating 80% and 81% accuracy, respectively, outperforming the current models.

The study [20] focuses mostly on the heterogeneous features of audio signals that differ based on the type of audio sentiment analysis performed. This research presents a deep neural network model based on utterances that recognizes audio features by combining CNN and LSTM. MFCC and other popular feature extraction algorithms are also used to identify homogenous features. Additionally, attention-based Bi-LSTM has been employed to combine the traits. The MOUD dataset from Spain

was used to test the model after it was trained on the MOSI dataset. The model performs 9.33% better than the state-of-the-art models. In [43], a newly developed framework for word-based Urdu speech sentiment analysis was created. Using MFCC, PLP, Spectral energy, and Chroma vector features, short-term audio characteristics are retrieved, and five mid-term features are then processed from these. The emotion of Urdu utterances is subsequently ascertained using these mid-term features as input. HMM and DTW are combined to obtain the ultimate view. A 600-word bespoke Urdu corpus with 97.1% accuracy is utilized to assess the model.

M. Sakurai and T. Kosaka [30] presented a new speech sentiment recognition method using the acoustic and lingual features of the data and achieved an accuracy of 82%. In their respective assessments of three universal speech representation variations for three sentiment analysis tasks and an emotion recognition task, Atmaja, B.T. and Sasou, A. [34] contributed. They used UniSpeech-SAT model to gain information about speech data and achieved 81% of performance. Researchers in [29] developed a sentiment-aware method for speech emotion recognition, which combined automatic speech recognition (ASR) and cross-entropy sentiment loss functions. They fine-tuned the model using the concordance correlation coefficient loss function (CCCL) to predict valence, arousal, and dominance. Results showed that integrating sentiment analysis with speech emotion recognition improved the accuracy of valence prediction. CCCL was found to be more effective than other loss functions [25]. Vimal et al. [32] implemented machine learning techniques to deal with Mel Frequency Cepstral coefficients (MFCC) and the energy of the speech signals. They classified their speech data into eight emotions. Using the RAVDESS dataset their Random Forest classifier achieved an accuracy of 88.54%. Moreover, they acknowledge the difficulty in processing speech data. Authors in their study [35] employed deep learning techniques such as CNN, GRU, and LSTM with MFCC and Mel Spectrograms for RAVDESS and TESS datasets gaining a 97.1% accuracy. The investigation of audio emotion recognition in this work [26] centers on the emotional responses elicited by non-musical sounds. To predict emotional aspects like arousal and valence, 76 parameters are examined using the International Affective Digital Sounds collection. Regression models are outperformed by machine learning techniques, such as shallow neural networks, which achieve prediction accuracy of up to 65.4%. For further research, the report recommends improving modeling techniques and improving datasets. An ensemble method named DSCA combining weak learners of Decision Tree, Linear SVC, CatBoost, and AdaBoost is proposed by Veni et al. [58]

Nonetheless, most Speech Sentiment Analysis (SSA) studies turned to speech-to-text (STT) methods to provide accurate sentiment classification from the textual contexts [8] [6] [11] [9]. Some of these studies include converting the input speech to text first and then making use of the sequence-based or NLP architectures to analyze the textual features. The summary of all the literature are provided in the Table 2.2

In summary, in both Bengali and English sentiment analysis tasks, the prevailing methods have combined basic machine learning techniques with state-of-the-art deep



Table 2.2: Literature Summary for Unimodal Methods

Paper	Dataset	Method	Performance(%)
[44]	IEMOCAP , RAVDESS	Attention-based GRU	81
[20]	MOSI	CNN and LSTM	Outperforms by 9.33%
[43]	A 600-word bespoke Urdu corpus	HMM and DTW	97.1
[30]	-	Aucoustic and linguistic features	82
[34]	CMU-MOSE	UniSpeech-SAT	81
[37]	MSP-Podcast	ASR and CCCL	70
[32]	RAVDESS	ML techniques	88.54
[35]	RAVDESS , TESS	CNN, GRU, and LSTM	97.1
[26]	International Affective Digital Sounds collection	Regression models and shallow NN	65.4
[58]	-	Decision Tree, Linear SVC, CatBoost, AdaBoost and Ensemble learner	-

learning approaches. Despite their advancements, these methods have not consistently outperformed traditional models. In Bengali sentiment analysis, datasets like SUBESCO, BangSER, and RAVDESS are commonly utilized, yet Ravdess’s inclusion of non-Bengali speech poses challenges for assessing real-life audio performance. To address this, our proposed method integrates three Bengali datasets (SUBESCO, BanglaSER and KBES), including real-life audio data from KBES. Moreover, we enhance traditional machine learning models with explainability, often overlooked in existing literature, to refine sentiment analysis systems and enhance interpretability and feature selection process. While recent studies have achieved impressive results with deep learning and transformer architectures, there’s untapped potential for traditional models to improve, aligning with our primary objective. Moreover, despite the success of advanced techniques, incorporating traditional models offers further refinement opportunities. By leveraging both traditional and modern approaches, our aim is to explore novel pathways for sentiment analysis, ensuring a comprehensive evaluation of audio data, especially in real-life scenarios. Additionally, challenges posed by unlabeled data in speech sentiment analysis remain underexplored. Many existing studies rely on converted image features from the audio, potentially impacting sentiment recognition accuracy. In our research, we address these issues by employing various feature extraction techniques and directly integrating them into our models, aiming for a comprehensive assessment of audio sentiment, considering nuances overlooked in previous approaches.

## 2.2 Multimodal Sentiment Analysis(MSA)

The last ten years have seen an increase in MSA research. Studies have specifically shown that it is useful for emotion recognition and sentiment prediction [50]. Transformer models and a bidirectional recurrent neural network-based module are used

in the most recent models. The task of combining the three modalities—text, audio, and video—and contextually choosing the most helpful ones falls to the attention-based module, which is the second module[50].

### 2.2.1 Bangla Multimodal Systems

While the multimodal models gained popularity by integrating modality-based features to refine the prediction of the models, most of the Bengali multimodal systems in the existing literature are based on Bengali memes. Hossain [40] presented a novel dataset for Bangla Meme Sentiment Analysis containing 4368 memes. Moreover, they carried out twenty-two experiments on unimodal and multimodal(image and text) models in which multimodal models outperformed the unimodal models with an accuracy of 64%. For visual modalities, they used pre-trained transfer learning architectures like VGG19, VGG16, ResNet50, and DenseNet121. For text-based modality various ML and DL techniques are employed. Feature concatenation is used for the fusion layer. An additional model for detecting hate speech from Bengali memes has been put forth by authors in [42]. They have combined the analysis of textual and visual data for hate speech detection by using Bi-LSTM/Conv-LSTM with word embeddings, ConvNets + pre-trained language models (such as multilingual BERT-cased/uncased, XLM-RoBERTa, and monolingual BanglaBERT). With an F1 score of 0.83, XLM-RoBERTa+ DenseNet-161 demonstrated the best performance in multimodal fusion. Expanding Bengali meme sentiment analysis systems To overcome the shortcomings of the study in [42], Elahi et al. [49] suggested an explainable multimodal strategy utilizing ResNet50 and BanglishBert. The final output is produced by passing the features through two linear layers by concatenating the feature vectors. With an F-1 score of 0.71, they succeeded. Bangla-BERT and XLM-R are utilized to fuse the features in another dataset for the same purpose that has been described in [41]. Furthermore, it has been suggested in [57] to use deep learning to identify the emotions in Bengali social media messages. They extracted visual features using transfer learning techniques including ResNet50, VGG16, and InceptionV3, and they used deep learning architectures like CNN and BiLSTM for textual content analysis. By utilizing Inceptionv3 and BiLSTM for feature fusion, they were able to achieve a f1 score of 77.5%.

### 2.2.2 Multimodal Sentiment Analysis Systems in Different Languages

Previously, the multimodal model with three modalities (text, audio, and visual) was built using HMMs in the work [4] for sentiment analysis from web opinions, and it obtained an F1 score of 0.553. Then, in [10], the authors suggested utilizing multiple kernel learning with a deep CNN to extract features for sentiment analysis and emotion recognition from textual and visual modalities. This feature selection technique effectively combines data from several modalities. Text, video, and audio comprised the multimodal data. The suggested strategy with feature selection marginally enhanced the performance of multimodal fusion without feature selection. In their study [23], Lu, Zhiyun, et al. developed an RNN with self-attention as the sentiment classifier, and by combining textual and audio information, it demonstrated the expected results. They employed the SWBD and IEMOCAP datasets,

with the SWBD dataset yielding the best accuracy of 70%. In order to predict the multi-dialect of speaker sentiment in the Arabic language in three dimensions—text, speech, and video—S. Al-Azani et al. [21] proposed an enhanced video analysis approach. The results indicate that the approach, which combines various methods of predicting the speaker’s feelings, may result in a more accurate prediction, with over 94% accuracy.

In another study [20] authors have proposed an utterance-based deep neural network learning method. They have used MFCC, Spectral acoustic features from the audio data used in LSTM models, and spectral graphs for CNN models. Moreover, they have used Bi-LSTM with an attention mechanism for the fusion of features. However, they achieved the highest 64% accuracy for the 2-class classification. Eric Chu and Deb Roy [12] leveraged the audio-visual modality for learning the emotional arcs in movies. They created a clustering technique to identify different emotional arc classifications. They used AlexNet as an image model and 96 bin Mel spectrograms as an audio model. However, in experiments, their proposed model could only achieve an accuracy of 0.652 while the F1 score was 0.741. In 2021, researchers presented a real-time feature extraction technique from audio-visual input using four deep neural networks. [33] In order to arrive at a final forecast using the REVDESS dataset, they combined the audio and visual emotion elements into a single stream and utilized an exponentially weighted moving average to gather evidence over time. Their suggested method outperformed the baselines, achieving an accuracy of 90.74%. They did acknowledge that their model did a worse job of predicting the affirmative statements.

In addition to this, a number of surveys on multimodal sentiment analysis have been conducted and published in the literature [50], [47], [61], and [51]. Together with transformer-based techniques [56][59] and attention-based mechanisms [52][60], translation-based studies [63][64][62] have been extremely popular recently. The summary of all the multimodal literature are provided in the Table 2.3

Table 2.3: Literature Summary for Multimodal Methods

<b>Paper</b>	<b>Dataset</b>	<b>Method</b>	<b>Modalities</b>	<b>Performance(%)</b>
[10]	-	Multiple kernel learning	Audio, video, and text.	-
[23]	IEMOCAP and SWBD	RNN with self-attention	Audio and text	70
[21]	-	Enhanced video analysis	Text, speech, and video	94
[20]	-	Utterance-based deep neural network	Audio and Image	64
[40]	MemeSen	DNN and RNN	Image and Text	64
[42]	Custom	XLM-RoBERTa+ DenseNet-161	Text and Image	0.83 F1 Score
[57]	Custom	BiLSTM and Inceptionv3	Image and Text	0.77 F1 score
[49]	MemeSen	ResNet50 and BanglishBert	Image and Text	0.71 F1 score
[4]	Youtube	HMM	Audio, Video, Text	0.553 F1 score
[12]	-	Clustering approach	Audio and Image	65
[33]	RAVDESS	Deep neural networks	Audio and Video	90.74

Table 2.4: Summary of the State-of-the-arts Speech Sentiment Analysis

Paper	Modalities	Language	Dataset	Method	Learning Type	Performance(%)
[45]	Audio	Bengali	SUBESCO, BanglaSER	Custom 1d CNN	Supervised	90
[48]	Audio	Bengali	SUBESCO, RAVDESS	MLP, Random Forest, Gradient Boosting, and SVM	Supervised	87.54
[43]	Audio	Urdu	A 600-word bespoke Urdu corpus	HMM and DTW	Supervised	97.1
[35]	Audio	English	RAVDESS and TESS	CNN, GRU, and LSTM	Supervised	97.1
[38]	Audio	Bengali	SUBESCO	Bi-LSTM	Supervised	83.33
[44]	Audio	English	IEMOCAP and RAVDESS	Attention-based GRU	Supervised	81
[34]	Audio	Universal	CMU-MOSE	UniSpeech-SAT	Self-Supervised	81
[37]	Audio and Text	English	MSP-Podcast	ASR and CCCL	Supervised	70
[40]	Image and Text	Bengali	MemeSen	DNN and RNN	Supervised	64
[27]	Audio	Bengali	RAVDESS, ABEG	SVM,KNN,AdaBoost, Logistic Regression, and XGBoost	Supervised	92
[33]	Audio and Video	English	RAVDESS	Deep neural networks	Supervised	90.74
[32]	Audio	English	RAVDESS	ML techniques	Supervised	88.54
[31]	Audio	Bengali	SUBESCO, RAVDESS	DCTFB	Supervised	86.86

[30]	Audio and Text	Japanese	JTES	Language model adaption	Supervised	82
[26]	Audio	English	International Affective Digital Sounds collection	Regression models and shallow NN	Supervised	65.4
[42]	Text and Image	Bengali	Custom	XLM-RoBERTa+ DenseNet-161	Supervised	0.83 F1 Score
[57]	Image and Text	Bengali	Custom	BiLSTM and Inceptionv3	Supervised	0.77 F1 score
[49]	Image and Text	Bengali	MemeSen	ResNet50 and BanglishBert	Supervised	0.71 F1 score
Proposed	Audio, Text and Image	Bengali	SUBESCO, BanglaSER, and KBS	Random Forest with SHAP feature selection	Semi-Supervised	Accuracy - 77 F-1 Score - 77

In a nutshell, as shown in the table 2.4 unimodal techniques are mostly experimented with and the majority of multimodal techniques for the Bengali language have integrated textual and visual elements; the potential for speech sentiment analysis to be multimodal has not yet been investigated. Furthermore, Bengali sentiment’s emotional acknowledgment hasn’t yet blossomed in contemporary writing. While multimodality has been investigated in other languages, mostly English, using various fusion techniques and approaches, audio features, text features, and Mel-Spectrogram features that are retrieved from the audio have not been employed to improve the interpretability of the systems. The most popular techniques for multimodalities are LSTM for text and audio and CNNs for images. In our approach, all three modalities—audio-acoustic characteristics, Mel spectrogram features, and transcribed text features—were used to implement speech sentiment analysis. Most of the existing studies didn’t use broad feature extraction techniques also while using mostly MFCC, ZCR, or only Spectral features. Additionally, semi-supervised learning is rarely used in the state of the art research. We developed the first semi-supervised multimodal system for Bengali Speech sentiment analysis that we are aware of, using sequential model LSTM for audio features, CNN for spectrogram data, and Transformers for textual features. We also integrated eight acoustic feature extraction techniques to merge features, the first in the literature. Moreover, our work has proposed a comprehensive analysis of performances using early fusion and late fusion techniques for multimodalities, which has not been explored in Bengali literature.

# Chapter 3

## Methodology

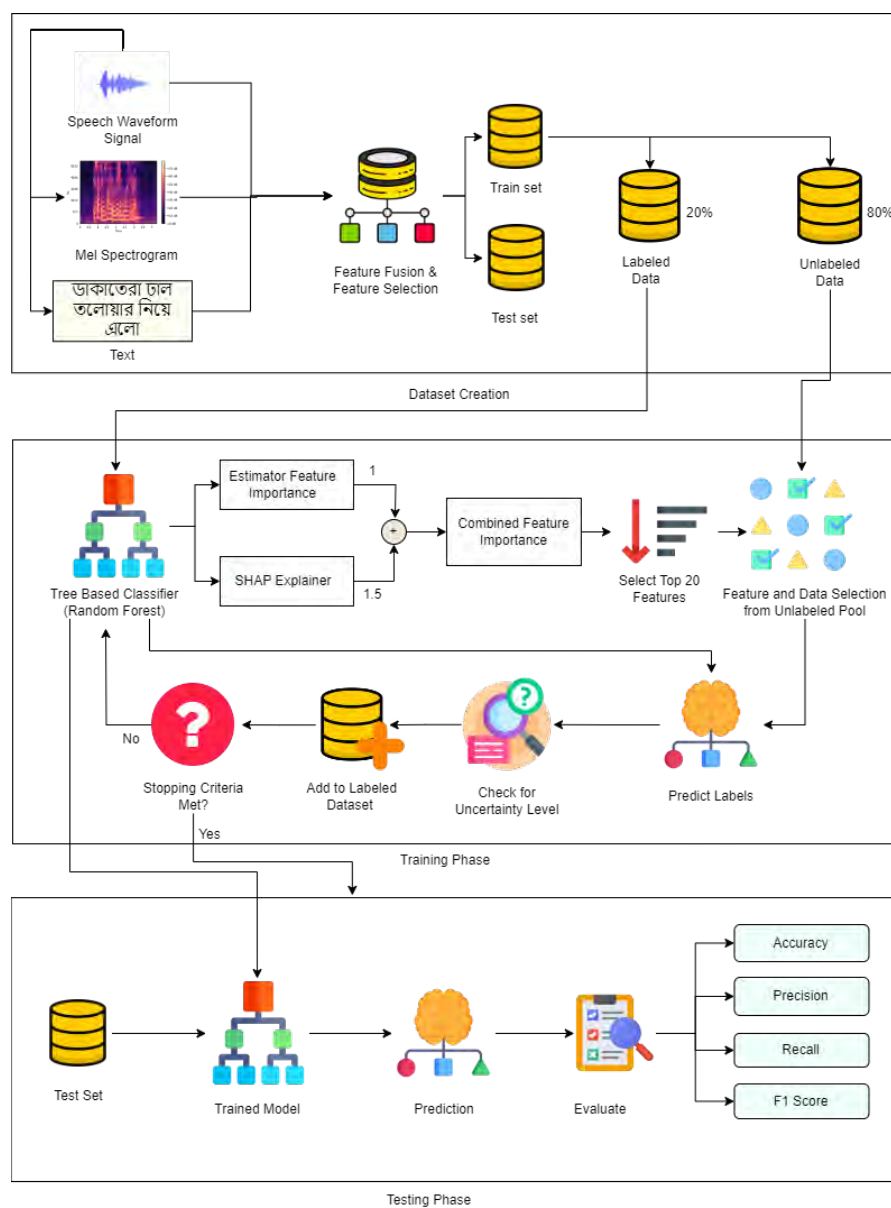


Figure 3.1: Proposed Early Fusion Semi-supervised Multimodal ML Approach with Iterative Feature Boosting using SHAP



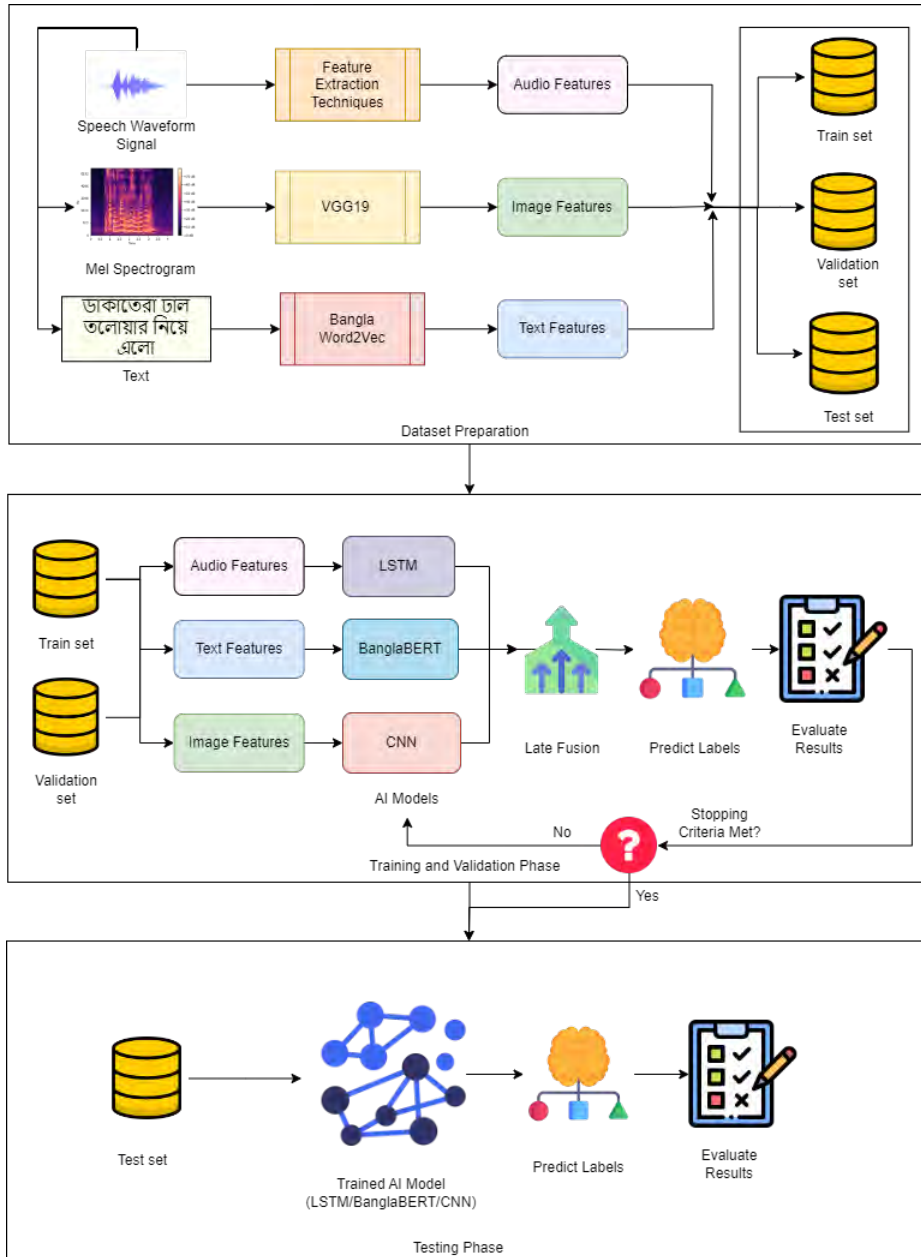


Figure 3.2: Proposed Late Fusion Multimodal Approach System Diagram

## 3.1 Dataset

The outcomes of models are significantly impacted by the appropriate dataset selection. The three datasets listed below were used in this study to train and evaluate our models in the Bengali language:

### 3.1.1 SUBESCO

There are 7000 audio speech files in this public audio-only emotional speech dataset [31] for the Bengali language, which is labeled with the following seven types of emotions: happy, sad, disgusted, fear, angry, neutral, and surprised. Ten male and ten female professional actors took part in the recording of ten statements representing seven target emotions. The collection, which was gender-balanced and

comprised trained native speakers, mimicked the seven emotions with an accuracy rate ranging from 71% to 80% in each recording of ten words. Each audio file has an average file duration of 4 seconds, and they are all available in the .wav format. The positive(Figure 3.3(a)), negative3.3(b)) and neutral(Figure 3.3(c)) class data samples visualization of this dataset is shown in Figure 3.3.

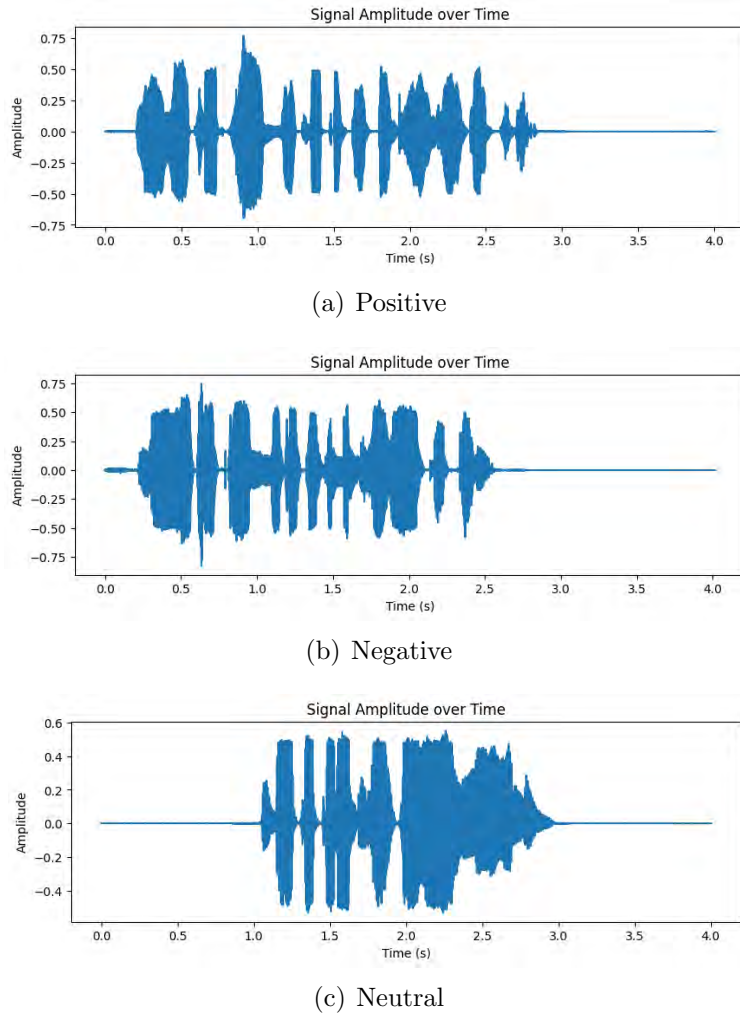
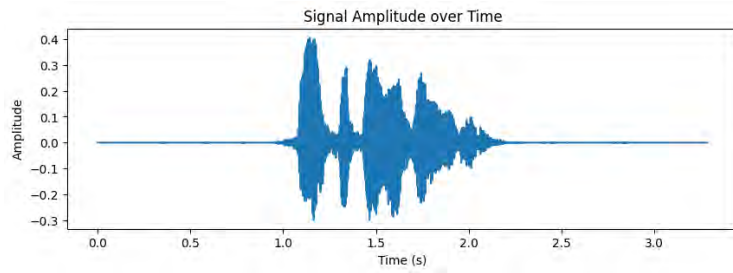


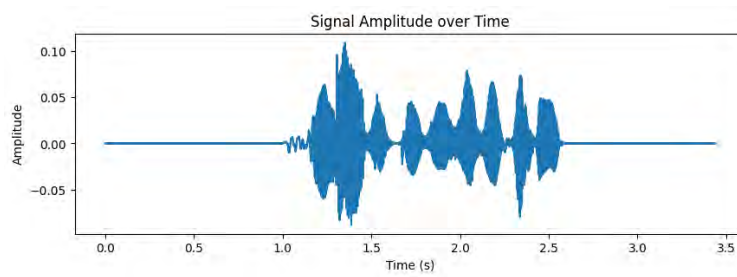
Figure 3.3: SUBESCO Dataset Data Samples

### 3.1.2 BanglaSER

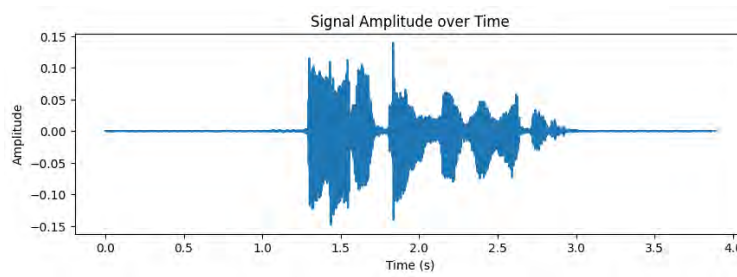
1467 Bangla speech-audio recordings featuring five distinct emotions—angry, joyful, neutral, sad, and surprised — created with the same proportion of male and female speakers make up this sentiment emotion detection dataset [36]. It includes speech-audio data from 34 speakers who participated, ranging in age from 19 to 47. The BanglaSER dataset is produced by recording speech audios on laptops and smartphones, with a fair distribution of male and female actors participating in each category and a balanced amount of recordings overall. The positive (Figure 3.4(a)), negative (Figure 3.4(b)), and neutral(Figure 3.4(c)) class data samples visualization of this dataset is shown in Figure 3.4.



(a) Positive



(b) Negative

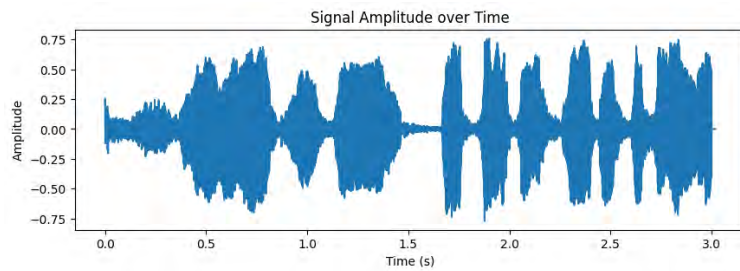


(c) Neutral

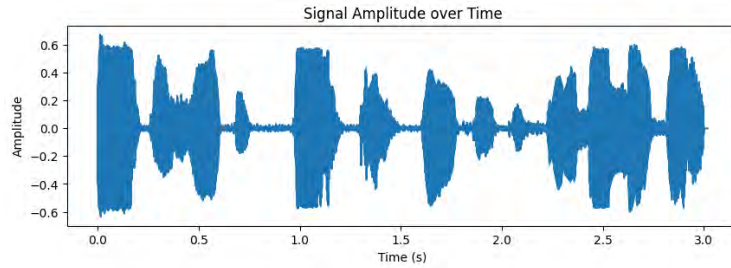
Figure 3.4: BanglaSER Dataset Data Samples

### 3.1.3 KBES (KUET Bangla Emotion Speech)

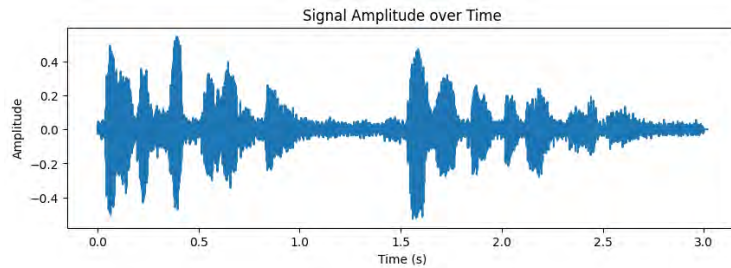
This dataset [46] contains 5 different emotions in speech which are Angry, Happy, Disgust, Neutral, and Sad. The dataset comprises 900 voice dialogs, or audio signals, from 35 performers, 15 of whom are male and 20 of whom are female, spanning a range of ages. There are two intensity levels for each sample of a given emotion: Low and High. The data are collected from YouTube and Facebook. The primary sources of the speech dialogue include dramas, TV shows, web series, and Bangla Telefilm. All the files are in .wav format. This dataset corresponds to the real-life data that will be used for speech sentiment analysis. The positive (Figure 3.5(a)), negative (Figure 3.5(b)), and neutral (Figure 3.5(c)) class data samples visualization of this dataset is shown in Figure 3.5



(a) Positive



(b) Negative



(c) Neutral

Figure 3.5: KBES Dataset Data Samples

### 3.1.4 Data Aggregation

For our main dataset, we have merged these three datasets (SUBESCO, BanglaSER, KBES) to build a new dataset. The reason behind merging is that the datasets were relatively small and had different types of speakers and tones. Then, based

on our understanding of human nature from these three datasets, we have classified the emotions and assigned a specific feeling to each one. We created our sentiment audio dataset by dividing the emotion label from each audio clip into three classes (Positive, Negative, and Neutral) according to the emotion that was received.

### **3.1.5 Dataset Distribution**

The emotions are categorized into each sentiment and are presented in the following table 3.1.

Table 3.1: Dataset Distribution Details

Sentiment	Emotion			Number of Samples per dataset				Total
	SUBESCO	BanglaSER	KBES	SUBESCO	BanglaSER	KBES	Total	
Positive	Happy	Joyful	Happy	1000	306	200	1506	2812
	Surprised	Surprised	-	1000	306	-	1306	
Negative	Angry	Angry	Angry	1000	306	200	1506	5312
	Disgusted	-	Disgust	1000	-	200	1200	
	Sad	Sad	Sad	1000	306	200	1506	
	Fear	-	-	1000	-	-	1000	
Neutral	Neutral	Neutral	Neutral	1000	242	100	1343	1342

From the table 3.1, we can see that three different datasets have some mostly similar emotions and some different emotions. The SUBESCO dataset has the same amount of data(1000) for every class. The number of neutral emotions is relatively less on both BanglaSER(242) and KBES(100) datasets and the number of negative emotions amount is much higher. That makes less amount of neutral data (1342) for the final dataset. We have merged the Anger, disgust, sadness, and fear emotions as negative emotions and happy and surprise as positive emotions. After totaling, we got 5312 audio data corresponding to Negative sentiment which is nearly 55.6% of the whole dataset as shown in the figure 3.6. Moreover, the amount of positive and neutral data are 2812 and 1342 respectively making them 30% and 14.3% of the whole dataset. As we can see there is an imbalance in the dataset from Figure 3.6, we have used techniques such as resampling and class weights to combat the imbalanced dataset which will be described in the later parts.

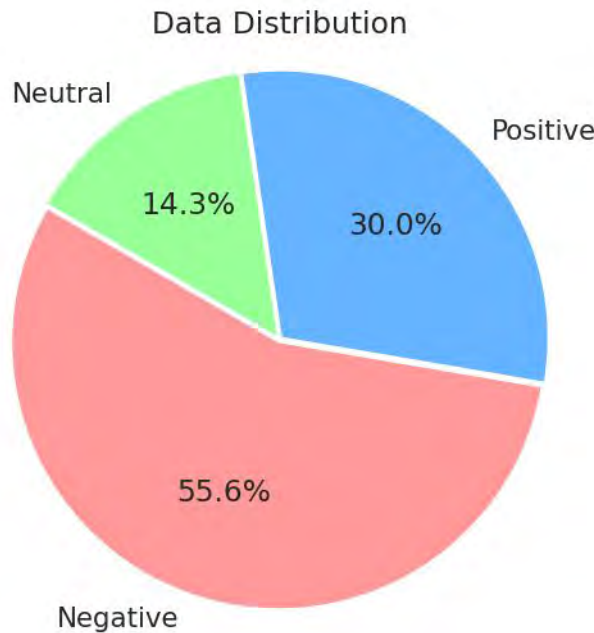


Figure 3.6: Data Distribution of Final Dataset

## 3.2 Data Augmentation

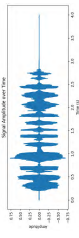
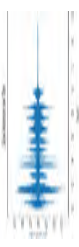



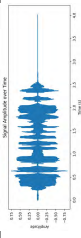




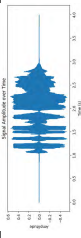




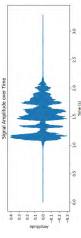




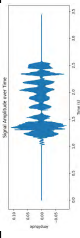

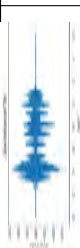

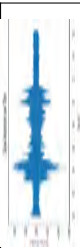
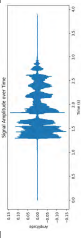




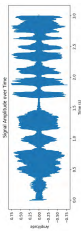




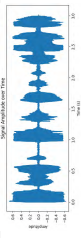

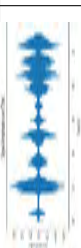


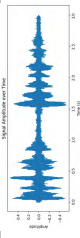


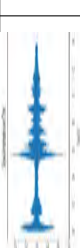

To enhance our models' capacity for generalization, we initially employed data augmentation techniques, turning the data into syntactic data. All of our data were in .wav format so we first converted them to mp3 format using the pydub library. As we can see, the audio in two of our datasets [61][62] was recorded in a controlled setting while taking many considerations into account. Real-world data is not always so pure and is not constrained by the surroundings. We altered the training dataset's raw audio by adding noise, pitch, and stretch to make the data more representative of real-world data.

To add the properties, we utilized the Librosa package for augmentation. We started by adding some random noise to our unprocessed data. Next, we added stretch factor

0.8, which lengthened the time by slowing down the data, after adding pitch factor 0.7, which we did by arbitrarily changing the frequency of the noisy data. From Table 3.2 we can see that for each dataset (SUBESCO, BanglaSER, KBES) we have added noise, pitch and stretch and merged them together to get the final data. Each step audio visualization is shown in the table. This process ensures that all of the data has variety and can perform well in real-life scenarios.



Table 3.2: Example of Data Augmentation for All Three Datasets

Dataset	Sentiment	Original Data	Added Noise	Added Pitch	Added Stretch	Final Data
SUBESCO	Positive					
	Negative					
	Neutral					
BanglaSER	Positive					
	Negative					
	Neutral					
KBES	Positive					
	Negative					
	Neutral					

### 3.3 Dataset Split

We divided our whole dataset into a 60:20:20 set for the Train:Validation: Test set. We shuffled data data before splitting. All the experiments were done with 80% of the data consisting of train and validation sets while 20% of data was kept aside for testing the model performance only.

### 3.4 Feature Extraction for Audio Modality

Extracting acoustic features from audio is one of the most important aspects of building a methodology for experiments as the model will work based on the features themselves. As we have used Speech audio data in our study, after data augmentation, we used different audio feature extraction techniques on the data to extract the acoustic features. The “Librosa” library has been used to do the feature extraction. The feature extraction details are given below -

#### 3.4.1 Mel frequency cepstral coefficients (MFCC)

Since the human ear has a nonlinear scale for audio perception, the MFCC aims to mathematically simulate the human ear. Mel frequency cepstral coefficients (MFCCs) of a signal are a small set of parameters (usually 10–20) that just specify the general spectral envelope shape [64]. We construct the pandas data frame of our training dataset, which consists of the 20 features, using librosa.feature.mfcc. MFCC extraction includes nine major steps. They are summarized as follows- The first step is Pre-emphasis which amplifies high frequency components. It is calculated as described in the equation 3.1-

$$x[n] = s[n] - \alpha s[n - 1] \quad (3.1)$$

Here:  $x[n]$  = Pre-emphasized signal

$s[n]$  = Input signal  $\alpha = 0.95 - 0.96$

The emphasized signal components are divided into overlapping frames where each frame has 50% overlap. The framed signal is calculated as shown in the equation 3.2. Then to reduce spectral leakage each frame is multiplied with a Hamming window as shown in the equation 3.3 which transformed into the frequency domain using Fast Fourier Transform (FFT) as equation 3.4. In the next step, power spectrum is computed as magnitude squared of the FFT shown in equation 3.5.

$$x_k[n] = x[n + kH], \quad k = 0, 1, 2, \dots, K - 1 \quad (3.2)$$

where ,  $H$  is the hop size (frame shift) and  $K$  is the total number of frames.

$$x_w[n] = x_k[n] \cdot w[n], \quad w[n] = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N - 1}\right) \quad (3.3)$$

Here,  $N$  is the frame length

$$X_w[k] = \sum_{n=0}^{N-1} x_w[n] \cdot e^{-j\frac{2\pi}{N}kn}, \quad k = 0, 1, 2, \dots, N - 1 \quad (3.4)$$

$$P[k] = |X_w[k]|^2 \quad (3.5)$$

Then the power spectrum is passed through mel filter get the energy in each mel frequency band which is described in the equation 3.6 and equation 3.7 and using this mel filter the Mel spectrum is obtained (equation 3.8). After that the logarithm of mel-spectrum is computed and then that value is transformed using the Discrete Cosine Transform to obtain the MFCCs shown in equation 3.9 and 3.10.

$$f_{\text{mel}} = 2595 \cdot \log_{10} \left( 1 + \frac{f}{700} \right) \quad (3.6)$$

$$f = 700 \cdot \left( 10^{\frac{f_{\text{mel}}}{2595}} - 1 \right) \quad (3.7)$$

$$S[m] = \sum_{k=0}^{N-1} P[k] \cdot H_m[k] \quad (3.8)$$

where  $H_m[k]$  are triangular filters spaced along the mel frequency axis.

$$\log S[m] = \log (S[m]) \quad (3.9)$$

$$C[n] = \sum_{m=0}^{M-1} \log S[m] \cdot \cos \left[ \frac{\pi n(2m+1)}{2M} \right], \quad n = 0, 1, 2, \dots, L-1 \quad (3.10)$$

where  $M$  is the number of mel filters and  $L$  is the number of desired cepstral coefficients (typically 12-13).

### 3.4.2 Zero Chroma Rate (ZCR)

The rate at which a signal's sign shifts from negative to positive or vice versa and creates an intermediate frequency is known as the zero crossing rate, or ZCR. Each frame's zero crossing is computed using the `librosa.zero_crossings()` function. For an input signal  $x[n]$  the zero chroma rate is calculated as shown in the equation 3.11 where the indicator function  $\mathbf{1}\{\cdot\}$  is expressed as equation 3.12

$$\text{ZCR} = \frac{1}{N-1} \sum_{n=1}^{N-1} \mathbf{1}\{x[n] \cdot x[n-1] < 0\} \quad (3.11)$$

Here,  $N$  = Frame length

$$\mathbf{1}\{x[n] \cdot x[n-1] < 0\} = \begin{cases} 1 & \text{if } x[n] \cdot x[n-1] < 0 \\ 0 & \text{otherwise} \end{cases} \quad (3.12)$$

### 3.4.3 Chroma Shift

To assess the tonal variations of the compressed audio signals, chroma shift has been employed. The `librosa.feature.chroma_stft()` function aids in identifying the audio's chords and determining harmonic similarity. [19]. From the input audio signal  $X(t,f)$  the chroma vector is calculated then the chroma shift is computed as shown in the equations 3.13 and 3.14.

$$c_k(t) = \sum_{f \in F_k} |X(t, f)| \quad (3.13)$$

$$c'_k(t) = c_{(k-s) \bmod 12}(t) \quad (3.14)$$

where  $s$  is the number of semitones to shift (positive for upward shifts and negative for downward shifts).

### 3.4.4 Root Mean Square (RMS)

To extract the features from the audio files, we utilized the `librosa.feature.rms()` function to perform Root Mean Square (RMS) which is commonly used to measure the energy level of the audio. The equation 3.15 shows the calculation for RMS.

$$\text{RMS} = \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} x[n]^2} \quad (3.15)$$

where:

- $x[n]$  is the value of the audio signal at the  $n$ -th sample.
- $N$  is the total number of samples in the signal or frame.

### 3.4.5 Spectral Centroid

Spectral centroid gives an idea of the power of the audio file where high centroid values indicate higher frequency contents. We have used the `librosa.feature.spectral_centroid()` function to compute the spectral centroid for the audio files. The calculation equation is shown in equation 3.16

$$C = \frac{\sum_{k=0}^{N-1} f[k] \cdot |X[k]|}{\sum_{k=0}^{N-1} |X[k]|} \quad (3.16)$$

Where,

- $f[k]$  is the frequency corresponding to the  $k$ -th bin.
- $X[k]$  is the magnitude of the spectrum at the  $k$ -th bin.
- $N$  is the total number of frequency bins.

### 3.4.6 Spectral Bandwidth

It gives details on the timbre qualities of audio and quantifies the signal's frequency spread. The equation 3.17 shows the calculation of Spectral bandwidth. The function `librosa.feature.spectral_bandwidth()` is utilized to extract data from audio.

$$B = \sqrt{\frac{\sum_{k=0}^{N-1} (f[k] - C)^2 \cdot |X[k]|}{\sum_{k=0}^{N-1} |X[k]|}} \quad (3.17)$$

where,

- $f[k]$  is the frequency corresponding to the  $k$ -th bin.
- $X[k]$  is the magnitude of the spectrum at the  $k$ -th bin.
- $C$  is the spectral centroid.
- $N$  is the total number of frequency bins.

### 3.4.7 Spectral Roll-off

The overall brightness or sharpness of the sound can be inferred from the spectral roll-off. A greater concentration of energy in lower frequencies is indicated by a higher spectral roll-off, whereas greater energy in higher frequencies is indicated by a smaller spectral roll-off. We computed the roll-off using `librosa.feature.spectral_rolloff()`. The roll-off is computed as shown in equation 3.18.

$$\sum_{k=0}^{k_{\text{rolloff}}} |X[k]|^2 = p \cdot \sum_{k=0}^{N-1} |X[k]|^2 \quad (3.18)$$

where,

- $X[k]$  is the magnitude of the spectrum at the  $k$ -th bin.
- $k_{\text{rolloff}}$  is the index of the frequency bin corresponding to the rolloff frequency.
- $p$  is the percentage of total spectral energy (typically  $p=0.85$  or 85%).

### 3.4.8 Spectral Flatness

The measurement of an audio signal's flatness, or peak, in the spectral domain, gives insight into the audio's tonal qualities and level of noise. The `librosa.feature.spectral_flatness()` method is used to determine this. An unpitched or noise-like sound is suggested by a generally flat spectrum, which is indicated by a high spectral flatness rating. The

flatness is calculated according to equation 3.19.

$$F = \frac{\left(\prod_{k=0}^{N-1} P[k]\right)^{\frac{1}{N}}}{\frac{1}{N} \sum_{k=0}^{N-1} P[k]} \quad (3.19)$$

where,  $P[k]$  is the power spectrum at the  $k$ -th bin.  $N$  is the total number of frequency bins.

After extracting all the features using the above techniques we have used the mean of all of the feature vectors except the MFCCs as they are more effective on the speech sentiment recognition and hold more audio features. Then, we stacked all the features together with the labels to build our final acoustic feature dataset.

## 3.5 Feature Generation for Image Modality

### 3.5.1 Melspectrogram Generation

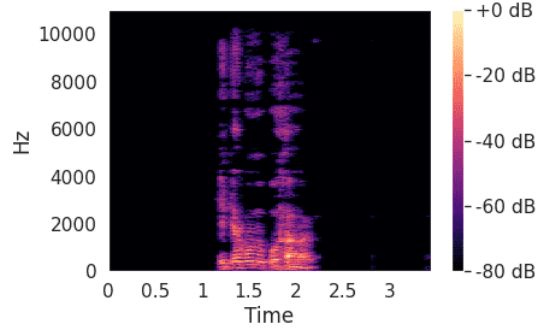
We have created Melspectrograms from every audio file for our multimodal models, which we will utilize as the Image Modality features. For all data of our dataset, we converted every audio file using the librosa library's Mel spectrogram approach. Each spectrogram was saved as a separate PNG image, to which the class labels were attached to create a CSV dataset file. We can see the generated mel spectrograms for the audio files in Figure 3.7

### 3.5.2 VGG19 Training

A pre-trained CNN architecture called VGG-19, which was introduced in 2014 [7], is said to yield excellent accuracies while processing huge picture datasets like ImageNet [3]. It was trained on a dataset of 1.2 million photos divided into around 100 categories. It lowers the size of the convolution filter and deepens the network with the aid of its three completely linked layers, sixteen convolution layers, and nineteen layers. Using feature maps, the convolution layers' 3 \* 3 design facilitates the identification of finer characteristics. All of the models have been categorized using the same set of hyperparameters. The ReLU activation function stacks each convolution layer. Faster computation and identification of the distinguishing characteristic of the mel-spectrogram pictures sorted according to attitudes are made possible by VGG-19.

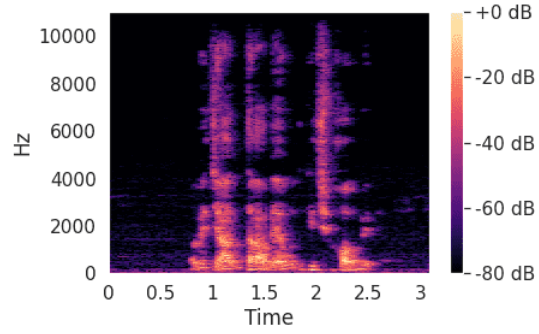
After Mel Spectrogram generation we transformed the images to size 224\*224 and fit them to VGG19 pretrained model without the top header layer(Fully connected layer and softmax) and zero number of classes. This training was done so that we could retrieve the features of the Mel Spectrograms from the trained model. For training the pre-trained model we have used the timm and poutyne library. Class weights were used in training the model. The model training details are given in table 3.3.

mel-spectrograms for sentiment positive dataset BanglaSER



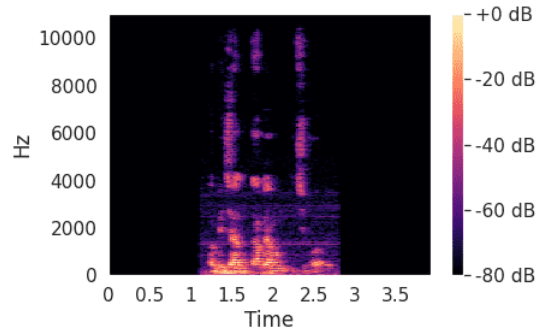
(a) Positive

mel-spectrograms for sentiment negative dataset BanglaSER



(b) Negative

mel-spectrograms for sentiment neutral dataset BanglaSER



(c) Neutral

Figure 3.7: Generated MelSpectrograms for Positive, Negative and Neutral Classes

Table 3.3: Hyperparameter Setting for Vgg19 Model Training

Parameter	Value
Input size	224*224
Epochs	20
Batch size	32
Learning Rate	1e-3
Loss	CrossEntropyLoss
Optimizer	Adam
Beta	0.9,0.999
Weight Decay	1e-5

After the training was completed we used the trained model to predict the features of the spectrograms and stored them in a CSV file along with the labels. The training details will be added to the appendix of the book.

## 3.6 Feature Generation for Text Modality

### 3.6.1 Audio to Text Transcribe

Text modality takes text as input and as we had speech audio files, to use them for text modality we had to transcribe the audio files to text. For Bangla speech, we have used a Speech recognition library with the parameter “bn-BD”. We transcribed all of the audio files and stored them in another CSV with the audio file path and labels for future reference.

Input	Transcribed Text	Label
audio1.wav	মৌমাছির চাক দেখে কুকুরটি ঘেউ ঘেউ করছে	Neutral
audio2.wav	ডাকাতেরা ঢাল তলোয়ার নিয়ে এলো	Negative
audio3.wav	বারোটা বেজে গেছে	Positive

Figure 3.8: Example of Transcribed Text from Audio Files

### 3.6.2 Generating Embedding Feature with BengaliWord2Vec

Machines don't work with raw texts, we need to feed the word embeddings to the machines to learn the representations of the text. That's why to use the texts on models we had to generate the word embedding of the texts. We had to first tokenize the texts and then generate embedding for each token to generate the word embedding. As shown in the figure 3.10 we generated the tokens and vector embeddings using the NLTKTokenizer and BengaliWord2Vec libraries from bnlp. The embedding size of the tokens is 100. After generating the embeddings these embeddings are used in an early fusion technique to merge with other modalities features to use as model input.



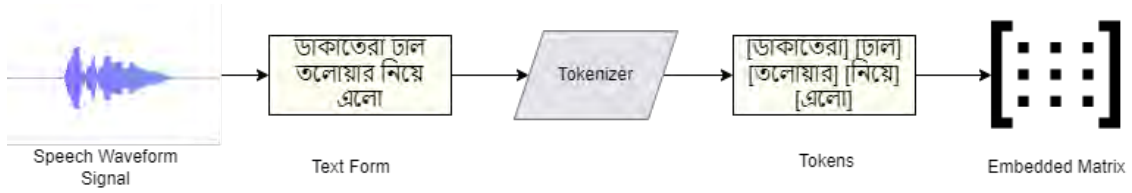


Figure 3.9: Process of Generating Embedding from Speech

### 3.6.3 Encoding for Transformer Input

For the late fusion technique, we have used the BanglaBERT transformer for the text modality. This transformer model takes token encoding as input along with the text token ID and attention mask. Therefore, we had to process the texts according to the desired inputs. For this process we used the BanglaBERT tokenizer from the pre-trained model of “csebuatnlp/banglabert”. The tokenization is done on the text with a max token length of 18 and we added special tokens needed for classification. Lastly, token lengths were padded to max length. We used the tokenizer’s parameter named `encode_plus` to get the integrated dictionary along with the extra parameters which are input id, token text id, and attention mask to feed it to the transformer model.

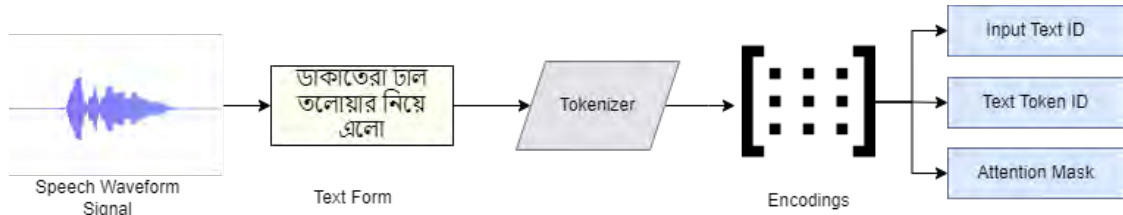


Figure 3.10: Process of Generating Encoding from Speech

## 3.7 Feature Selection

With the use of the 20% validation data, we ran an ablation experiment to choose the best feature selection method for our data. We experimented with base Random forest and Adaboost models as validators. The seven feature selection techniques we experimented with are -

- Recursive feature elimination: It operates by recursively deleting attributes and creating a model from the remaining attributes. It determines which attributes contribute most to the prediction of the target characteristic by using the model’s accuracy.
- Tree based methods: In order to pick features, tree-based techniques like Random Forest and Gradient Boosting evaluate each feature’s significance during the tree-building process.
- Principal Component analysis: The dimensionality reduction method known as principal component analysis (PCA) converts the initial features into a new set of uncorrelated variables known as principal components.

- Correlation-based feature selection: The association between every attribute and the ultimate variable, as well as the correlation between various features, is measured using correlation-based feature selection.[69]
- Mutual information: The statistical reliance between two variables is measured by mutual information. It measures the amount of information that a single feature gives about the target variable in the context of feature selection. [69]
- Sequential feature selection: Sequential Feature Selection is a technique that employs a machine learning model to assess the performance of several feature subsets by combining them.
- Recursive feature elimination: RFE is a straightforward but efficient feature selection technique that is used iteratively to remove the least important features from a model until the required number of features is achieved. [67]

As the AdaBoost model was the best-performing model in terms of the test set in the study [55] we selected the feature selection method which performed best with the AdaBoost model. The detailed results regarding the experiments can be found in the section Results. From our experiments, we could see that Correlation-based feature selection performed best with the validation dataset which is why it was chosen to be the technique we used further in our methods. This technique is used to find a subset of features that are highly correlated with each other in terms of the target variable and removes the redundancy of the features.

## 3.8 Re-sampling and Normalization

**Re-Sampling:** Our dataset featured an uneven distribution of data across the three groups because it was a combination of three separate datasets as shown in the dataset section Figure. To balance the dataset, we, therefore, employed the oversampling technique. To resample the dataset, SMOTE from the library `imblearn.over` sampling was utilized.

**Normalization:** Next to resampling we have scaled to standardize our data. Our dataset has been transformed and normalization has been carried out using `StandardScaler()`.

## 3.9 Multimodal Models

### 3.9.1 Early Fusion Model

#### Enhanced Machine Learning Model

As in the early fusion technique, we fuse the features into one dataset and run them on a single classifier model, we have also experimented with our proposed enhanced ML model which was used for unimodal classification. Three experiments were done for three combinations of modalities which are - Audio-Text, Audio-Image, and Audio-Text-Image. Feature selection was done on whole features irrespective of modalities to observe the impact of the dimensionality of features on the model.

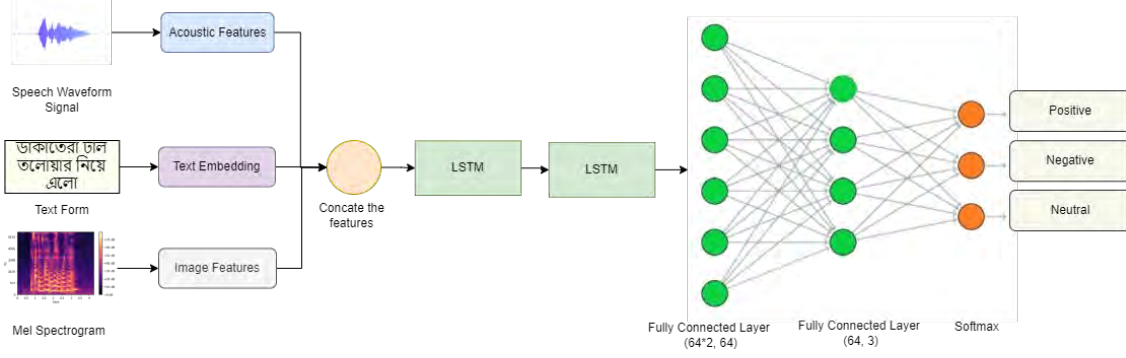


Figure 3.11: Proposed Early Fusion LSTM Model

We also compared the results with deep learning model performances. We selected 50 features based on the correlation and 20 features based on the weighted feature importance. Only the Random Forest model was used as the base model for the multimodal technique as it performed better for the unimodal systems. The detailed results are provided in the Results section.

### Proposed Random Forest Model

The supervised machine learning model known as the random forest, or random decision forest, employs a variety of learning techniques to perform tasks such as regression and classification. To improve the anticipated accuracy of the dataset, it applies multiple decision trees to various subsets of the input dataset and averages the outcomes. The input dataset is divided up into random subsets. For every data subset, a decision tree is built with the aid of the Gini impurity. The parent node is split further if the total GINI impurity of the split sub-tree is less than the GINI impurity of the parent node. Finally, the final output is chosen using a bagging process depending on the majority of votes. We used a Random forest model from sklearn with 50 estimators and gini impurity.

$$I = G_{\text{parent}} - G_{\text{split1}} - G_{\text{split2}} \quad (3.20)$$

$$G = \sum_{i=1}^C p_i * (1 - p_i) \quad (3.21)$$

$$EuclideanDistance(x, y) = \sqrt{\sum_{i=1}^n (x_i^2 - y_i^2)^2} \quad (3.22)$$

where,

$(x_i, y_i)$  — x,y coordinates of data samples

$n$  — total number of data samples

Here, C = Total Class number G = Gini Impurity I = Intensity

### Top Feature Selection and Weighting Strategy

SHAP (SHapley Additive exPlanations) is used for explaining machine learning output and it is based on game theory. The local and global feature importances of

the model for the predictions can be retrieved from SHAP. In our model, we have used the SHAP to generate the tree explainer and then generate the SHAP values. Then, we calculate the feature weights from the estimators. The Random Forest model `feature_importances` parameter provides us with the data, but we calculate it using individual estimator feature importance and weight for the AdaBoost model. In the next step, we add both of the values by multiplying them with a custom weight. For, SHAP importances we use 1.5 and for estimators, we keep it as it is. We experimented with taking the weight and only SHAP or only estimator feature importances. Those results are shown in the results section. As the best result is achieved by weighted strategy we selected it. Next, we summed the weights sorted all the feature's importance, and selected the top 20 features based on summed feature importance.

## Model Description Step by Step

---

### Algorithm 1 Semi-Supervised Learning with Feature Selection with ML Model

---

**Require:** Training dataset  $D$  with  $n$  data points, labeled pool  $L$  (20% of  $D$ ), unlabeled pool  $U$  (80% of  $D$ )

**Ensure:** Trained model and evaluation results

- 1: Divide the training dataset  $D$  into labeled dataset  $L$  (20%) and unlabeled dataset  $U$  (80%).
  - 2: **repeat**
  - 3:   Train the base model  $M$  with the labeled pool  $L$ .
  - 4:   Generate the model's feature importance using SHAP and model weights.
  - 5:   Select the top 20 features.
  - 6:   Select random data points from the unlabeled data  $U$ .
  - 7:   Select the top 20 features from the selected data points.
  - 8:   Generate predictions from the trained model  $M$  for the selected data points.
  - 9:   **for** each selected data point  $x \in U$  **do**
  - 10:     **if** uncertainty level is low **then**
  - 11:       Add the predicted label to  $x$  and move  $x$  from  $U$  to  $L$ .
  - 12:     **end if**
  - 13:   **end for**
  - 14: **until** stopping criteria are met (manually decided accuracy level by trial and error)
  - 15: Test the final model  $M$  with the test set.
  - 16: Generate and evaluate the results.
- 

## Deep Learning Model

### Long Short Term Memory (LSTM)

In 1997, Hochreiter and Schmidhub first suggested that LSTM was an RNN variation [2]. LSTM typically only processes data streams in a forward fashion. These models outperformed ordinary RNN networks in the case of consecutive audio recordings, resolving classical problems like long-term reliance and short-term memory (vanishing gradient problem). An LSTM unit composed of these three gates and a memory

cell, also called an LSTM cell, can be thought of as a layer of neurons in a standard feedforward neural network, where each neuron has a hidden layer and a present state. The input gate, forget gate and output gate are the three different kinds of gates found in an LSTM. Information enters the memory cell under the direction of the input gate. Information exiting the memory cell is regulated by the forget gate. Information exiting the LSTM and going into the output is managed by the output gate. The LSTM can preserve long-term dependencies in the input data by using the gates to selectively forget or keep information from earlier time steps. In the figure 3.12, we can see the structure of LSTM with its equations. Here,  $h$  denotes the hidden state,  $c$  denotes the cell state and  $o$  denotes the output.

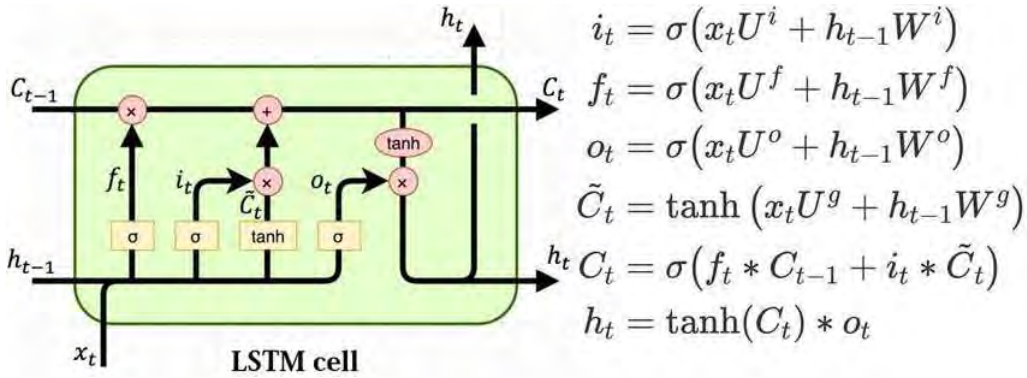


Figure 3.12: LSTM Structure with Equations

### 3.9.2 Late Fusion Models

#### Model used for Audio Features - LSTM

For all of our multimodal models, we have used the LSTM model for processing the audio features. Using the Pytorch package, a 3-layer LSTM neural network model with 15 epochs per layer, a learning rate of 0.0001, and a batch size of 32 has been constructed. The Adam optimizer and the ReLU and CrossEntropyLoss functions have been incorporated into the LSTM model. The input of the LSTM model was the audio features and the output varied according to the different models. The decision of which layer of the final model output to use depended on the intended models, which will be described later.

#### Model used for Image Features - Convolutional Neural Network (CNN)

Research on computer vision techniques for deep learning models is now the most popular area. CNNs are widely used for classification problems and typically consist of several convolution layers and fully connected layers. Convolutional neural networks are made to take advantage of the fact that images are 2D, which lets them learn hierarchical representations of image features through convolutional layers. The Pooling layer reduces the spatial size of the Convolved Feature. Max pooling finds the max value from the kernel and replaces the value. The dropout layer is a mask that eliminates certain neurons' contributions to the layer below while leaving all

other neurons intact.

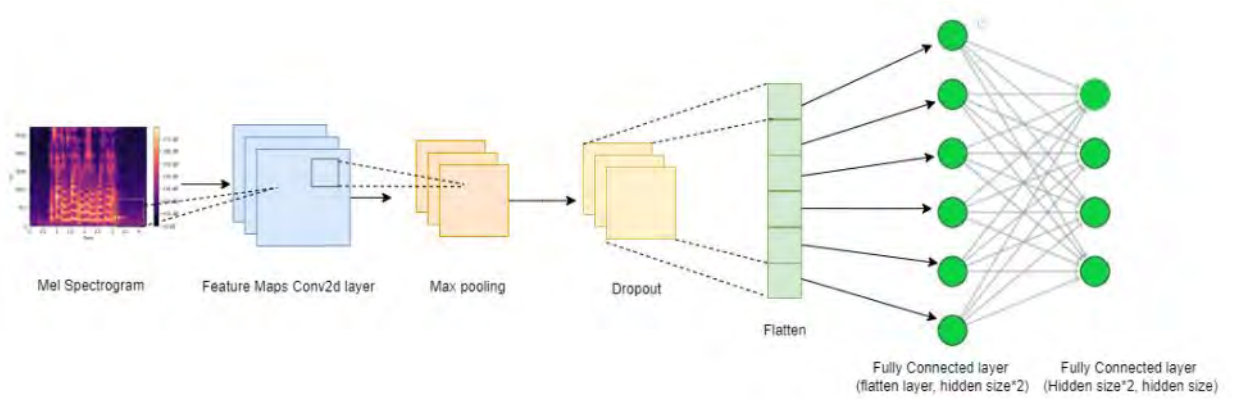


Figure 3.13: Custom CNN for generating Visual Features

In this work, for image modality, we have concentrated on classifying speech sentiments using CNN models by converting audio signals into spectrogram images. After using the features extracted from VGG19 models, we have used them in a custom CNN model. For the custom CNN model, we have built a 2d convolutional layer using the PyTorch library as shown in the figure 3.13. A kernel size of 3 is used for the model along with a maxpool layer and that is followed by a dropout of 0.3. The input size of the model is 224\*224 with 3 channels (RGB). The dropout layer output is then flattened and passed through two fully connected layers (hidden\_size\*2 and hidden\_size) to generate the final hidden state.

### Model used for Text Features - BanglaBERT

Transformer performs exceptionally well with sequential data, and as we have used text features we selected a pre-trained BERT (Bidirectional Encoder Representations from Transformers) language model. However, as our study works with the Bengali language we have opted for the BERT version BanglaBERT which has been pre-trained for the Bengali language [69]. This model is a pre-trained ELECTRA discriminator using the Replaced Token Detection (RTD) goal. This was carried out on a layer Transformer encoder (256 batch size for 2.5M steps) with 768 embedding size, 768 hidden size, 12 attention heads, 3072 feed-forward size, generator-to-discriminator ratio 1/3, and 110M parameters.[69]. The basic structure of BERT model is shown in the figure 3.14

As input for the model, we used the encodings done with the same model's tokenizer. We also used the attention mask to pay attention in the training process. For output, the output layer with hidden states has been used in different models.

### 3.9.3 Model Training Details

#### Early Fusion Training For all Modality Combinations

As the name implies, we must fuse the features for the early fusion model before feeding them into the model network. As a result, we first combine the features

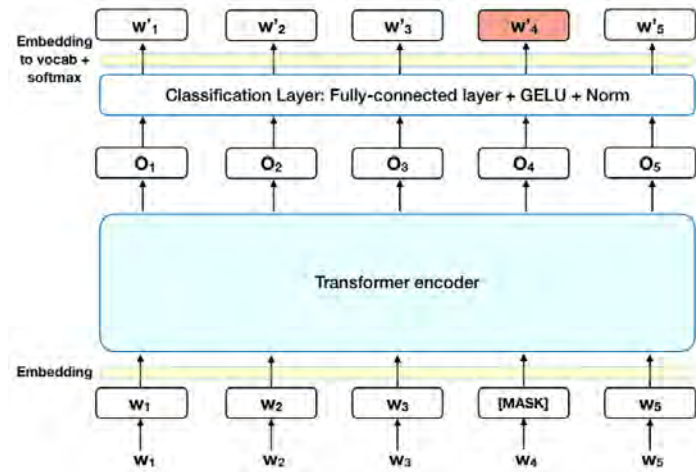


Figure 3.14: BERT structure (Image taken from [68])

Table 3.4: Hyperparameters of Early Fusion Model Training

Parameter	Value
Input size	Audio, Text and Image : 151 Audio and Text: 138 Audio and Image: 51
Output size	Number of classes : 3
Epochs	50
Batch size	32
Hidden Size	64
Activation Function	ReLU
Learning Rate	0.001
Loss	CrossEntropyLoss
Optimizer	Adam

of different modalities into a single dataset, scale the resulting dataset, and determine the class weight. The same model is used for three experiments and different modalities - Audio-Text, Audio-Image, and Audio-Text-Image. We perform feature selection using the correlation for the audio data. Next, since LSTM is a sequential model that performs well with acoustic, visual, and textual information, we employ it for all of our early fusion models. The feature set is given as the input, and a three-layer LSTM model is constructed. The model is trained for 50 epochs with a batch size of 32, or until the halting condition is satisfied. We utilize a patience level of 3 as the halting criterion to check the continuous reduction of validation loss. We use the last hidden state from the LSTM model output and pass it through two fully connected layers of shape fc1 (hidden size\*2, hidden size) and fc2 (hidden size, output size). Then our model is validated with the 20% validation dataset. Next, the validation evaluation is done for the model. We saved each of the epochs and used the best model state for testing our test dataset.

The model details are given in the table 3.4 -

After the completion of training and validation, the best model state is used to test the model for unseen test data and generate the evaluation metrics.

Table 3.5: Hyperparameters of Audio Text Modal Late Fusion Model Training

Parameter	Value
Input size	Audio: 38 Text: 18
Output size	Number of classes : 3
Epochs	50
Batch size	32
Hidden Size	64
Activation Function	ReLU
Learning Rate	0.0001
Loss	CrossEntropyLoss
Optimizer	Adam

### Late Fusion - Audio and Text

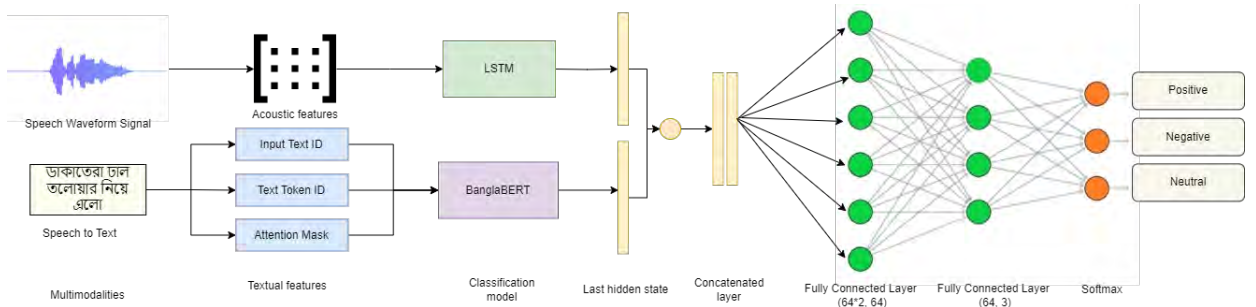


Figure 3.15: Proposed Late Fusion Model for Audio and Text

For the multimodal model with Audio and Text, we first store each audio feature file and the corresponding text feature file separately. Feature selection is performed for the audio files. Next, we build the models. LSTM is used for audio and BanglaBERT for text. The last hidden state output from the LSTM model is taken and the hidden state output from the Bert model is taken as the features of fusion. We use the concatenation method as a late fusion technique. After the concatenation, the combined outputs are passed through two fully connected layers and the ReLU activation function to generate the final output. The model is trained for 50 epochs. Using an early stopping patience value of 3, validation loss is monitored, and training of the model is stopped when validation loss becomes stagnant. The data was provided as a batch size of 32 and Adam optimizer is used for the same. The other details of the model hyperparameters can be found in Table 3.5. Also, the full system diagram can be found in Figure 3.15.

### Late Fusion - Audio and Image

As illustrated in the figure 3.16, first the audio and image features are stored separately and sent to their respective models for training. As described earlier, LSTM for audio and 2D convolution neural network is used for the image data. Feature selection is performed for audio data where we selected 20 features from the total features. The images for images were transformed by resizing and normalizing them



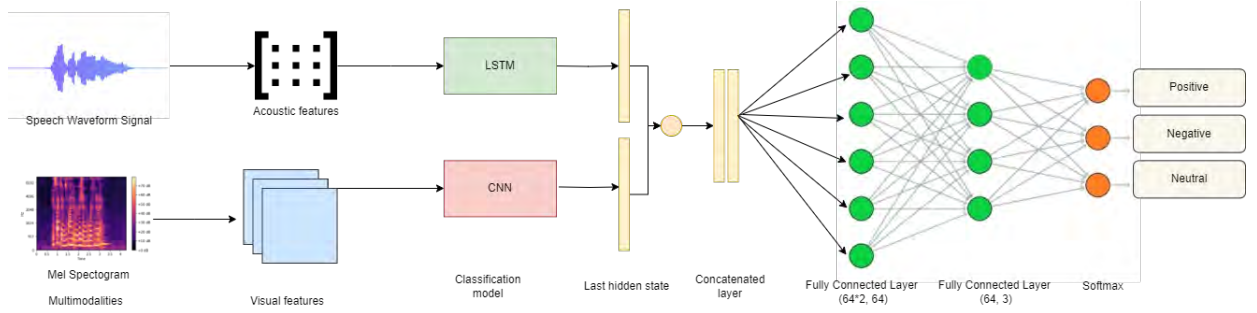


Figure 3.16: Proposed Late Fusion Model for Audio and Image

Table 3.6: Hyperparameters of Audio Image Modal Late Fusion Model Training

Parameter	Value
Input size	Audio: 38 Image: 224*224*3
Output size	Number of classes : 3
Epochs	50
Batch size	32
Hidden Size	64
Activation Function	ReLU
Learning Rate	0.0001
Loss	CrossEntropyLoss
Optimizer	Adam

before training. The images are provided as input to a convolution layer and passed through a maxpool layer onwards. A dropout of 0.3 is added to the output. Then this output is concatenated with the audio output which is received as the last hidden state of the LSTM model. These combined features are passed through two fully connected layers followed by the ReLU activation function. After that the loss is calculated using the CrossEntropyLoss function and the Adam optimizer is employed. After that the model is validated with the validation set and validation evaluation metrics are calculated. The training is done in 32 batches for 50 epochs with an early stopping technique. If the validation loss is not decreased for straight 3 rounds the training is stopped. After completing the training and model is tested with the test set to generate evaluation results. The details of the model are given in the table 3.6

### Late Fusion - Audio, Image, and Text

For all three modalities, we have combined the three models. First separate datasets were prepared for those models. 20 audio features were selected using the correlation technique and textual and visual features were kept as it is. LSTM, CNN, and BanglaBERT models were used separately for acoustic, visual, and textual features respectively. The model details were kept the same as the previous ones. Taking the last hidden state of all the models we combined them using the concatenation technique. The fused features are then passed through a fully connected layer and ReLU activation function followed by a 0.5 dropout. Finally, they are passed through

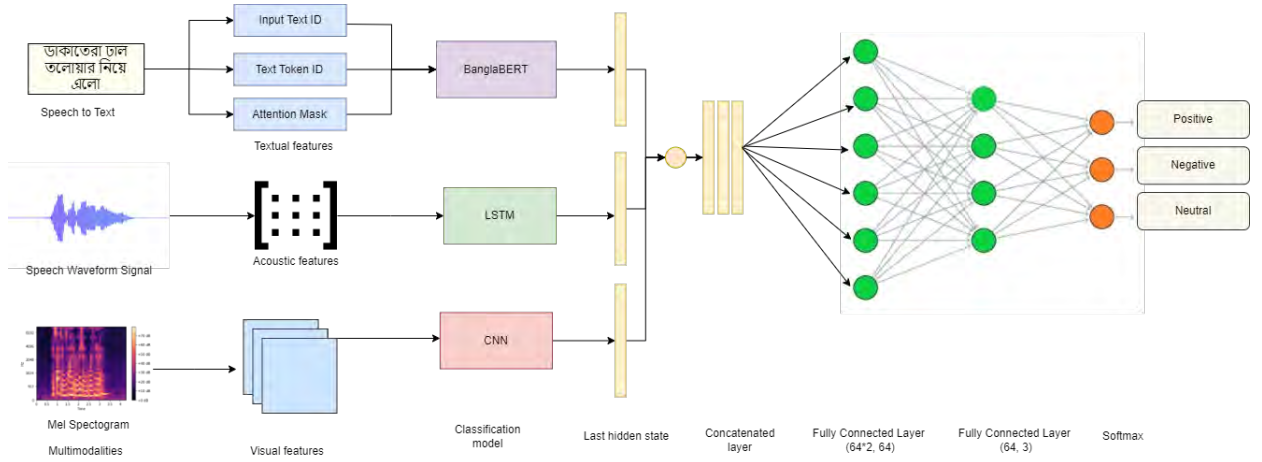


Figure 3.17: Proposed Late Fusion Model for Audio, Text and Image

Table 3.7: Hyperparameters of Audio, Text, and Image Modal Late Fusion Model Training

Parameter	Value
Input size	Audio: 38 Image: 224*224*3 Text: 18
Output size	Number of classes : 3
Epochs	50
Batch size	32
Hidden Size	64
Activation Function	ReLU
Learning Rate	0.0001
Loss	CrossEntropyLoss
Optimizer	Adam

another fully connected output layer to generate the final output of 3 classes of probabilities. Keeping the same pattern across all the models the training with validation is done in 32 batches for 50 epochs until stopping criteria are met. The best state of the model is loaded and the test is done with the test set. The whole process is illustrated in the Figure 3.17. The model details are shown in the following table 3.7 -

### 3.10 Unimodal Models

For the unimodal models, we used the acoustic feature dataset which we built according to the previous section. We experimented with multiple models and tested them with a custom test dataset as described in [55] and finally came to the conclusion that the traditional machine learning models - Random Forest and AdaBoost perform best with the acoustic features. We used them as our base model and enhanced them to build a model that can iteratively learn from predictions and select the most informative features using explainable AI SHAP.

### 3.10.1 Random Forest Model

For the Random Forest, the same model is used as described in the multimodal section. For the input, only acoustic features are used after feature selection.

### 3.10.2 AdaBoost Model

An embeddable learning model called AdaBoost (Adaptive Boosting) was introduced to increase the prediction ability of weak learners. To tackle binary classification issues, it builds a strong learner by combining the errors of weak learners. In order to create multiple decision stumps, AdaBoost divides and splits the samples into two subsets of a single feature, predicts the output based on these subsets, and then computes the decisions using GINI impurity. Because they focus exclusively on one aspect, these are referred to as weak learners. However, in a real-world situation, a decision is dependent on a number of factors, which is where ensemble learning enters the picture. To begin with, each example has the same weight.

$$w = 1/N \tag{3.23}$$

where,

- $w$  — data sample weight
- $N$  — the total number of data samples

After that, the AdaBoost method learns through the errors made by the weak learners in its predecessor and generates a new decision tree that places greater weight on incorrectly categorized features until data samples are predicted correctly. Choice In a Random Forest, stumps resemble trees but are not "fully grown." They have two leaves and one node. AdaBoost does not use trees, but rather a forest of these stumps.

$$\alpha_t = \frac{1}{2} \ln \frac{(1 - totalError)}{totalError} \tag{3.24}$$

where,

$\alpha$  = The degree to which the stump influences the classification

$$totalError = \frac{total\ number\ of\ misclassified\ data\ points}{total\ number\ of\ data\ points} \tag{3.25}$$

The weight value is updated using alpha. It is shown as follows-

$$w_i = w_{i-1} * e^{\pm\alpha} \tag{3.26}$$

Here, the value of alpha will be positive for correctly classified data and negative for misclassified data. In our model, we have used 50 estimators with a learning rate of 0.5 and used DecisionTreeClassifier as the base model. The parameters were decided by running a grid search on the parameters for the 20% dataset. Those details will be added to the appendix section.

# Chapter 4

## Results & Discussions

In this section of the thesis, we will describe the findings and conclusions of our research. The result section is broadly divided into three sections - Feature selection, Unimodal System Results, and Multimodal system results. First, we will describe the experimental results and selection of the feature selection method. Then we will dive into main system results with unimodal and multimodal methods.

### 4.1 Performance Metrics

This investigation employed a range of performance metrics to investigate why machine learning models could exhibit strong performance when measured by one evaluation metric but poor performance when measured by another. Accuracy, Precision, Recall, and F1-Score were the primary performance evaluation criteria employed in this study. The figure 4.1 demonstrates the confusion matrix for the evaluation metrics

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

Figure 4.1: Confusion Matrix for Evaluation Metrics

### 4.1.1 Accuracy

According to the equation 4.1, accuracy is calculated by dividing the total number of correct predictions by the total number of data samples in the dataset. The ratio of true positives (TP) to true negatives (TN) to the total number of samples is used to compute it-

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.1)$$

### 4.1.2 Precision

According to the equation 4.2-, precision is calculated by dividing the total number of correctly predicted positive outcomes by the total number of positive forecasts.

$$Precision = \frac{TP}{TP + FP} \quad (4.2)$$

### 4.1.3 Recall

The recall is defined as the total number of accurate positive predictions divided by the total number of actual positive predictions as shown in the equation 4.3. It's computed as the ratio of true positives (TP) to the total of false negatives (FN). -

$$Recall = \frac{TP}{TP + FN} \quad (4.3)$$

### 4.1.4 F-1 Score

As shown in the equation 4.4, F1-Score is the harmonic mean of precision and recall. Because it penalizes strong negative values of either component, the F1 Score is helpful when attempting to strike a compromise between good recall and high precision.-

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \quad (4.4)$$

In all of the equations -

TP = True Positive,

TN = True Negative,

FP = False Positive,

FN = False Negative

While precision and recall concentrate on the caliber of positive and negative predictions, respectively, accuracy assesses the overall correctness of the model's predictions. Because it strikes a compromise between recall and precision, the F1 Score is a more thorough metric for assessing classification models.

## 4.2 Experimental Setup

All the experiments are done with Google Colab T4 GPU with 8GB CPU RAM and 16GB GPU RAM. To explore and select the initial dataset we used a LENOVO Ideapad 310 laptop with Intel(R) Core(TM) i7-7500U CPU @ 2.70GHz 2.90 GHz equipped with 8 GB RAM which operates with Windows 10 system. The PyTorch and Poutyne library is used for Deep learning models and Scikit learn. Library for tree-based models. PyTorch Image Models (timm) are also used for the training process. Other than that Pandas, NumPy, Librosa, seaborn, matplotlib, pydub, Speech recognizer, SHAP, SHAPforAdaBoost and other mentioned libraries are also used.

## 4.3 Ablation Study

### 4.3.1 Feature Selection

We used seven feature selection techniques with validation to identify the most accurate feature selection technique. The table shows the outcomes that the approaches produced. The correlation-based feature selection method produced the best accuracy, 58.5% over the 20% dataset, as can be seen in the table 4.1. While every method demonstrated an accuracy of more than 50%, recursive feature reduction yielded the lowest accuracy, at 53.7%. With 57.4% accuracy, Principal Component Analysis (PCA) also fared well.

Table 4.1: Results of Feature Selection Test

Method	Accuracy
Recursive feature elimination	0.5372
Tree-based methods	0.5797
Principal Component analysis	0.5744
Correlation-based feature selection	<b>0.5851</b>
Mutual information	0.5478
Sequential feature selection	0.5425
Recursive feature elimination	0.5531

### 4.3.2 Results of Unimodal Systems

We have implemented an enhanced semi-supervised learning system with iterative feature boosting using the Random Forest and AdaBoost models as based and experimented with different setups. In this section, we will describe the experiments and results of the unimodal systems.

#### Basic Models

For the first experiment, we assessed the performance of the traditional AdaBoost and Random Forest model with all the labeled features. This is used to compare the performance of the proposed system with unlabelled data and fewer features. From the table 4.2 , we can see that with 50 features, the Random Forest model

achieves a weighted accuracy of 82% while precision, recall, and F1 scores are also the same. Adaboost performs poorly compared to Random Forest with 62% accuracy. It is seen from the confusion matrices in figure 4.3 and 4.2 that, AdaBoost fails to predict the correct sentiment for Negative speech data by falsely classifying 461 data to positive class where only 230 negative data are misclassified as positive by the Random forest model.

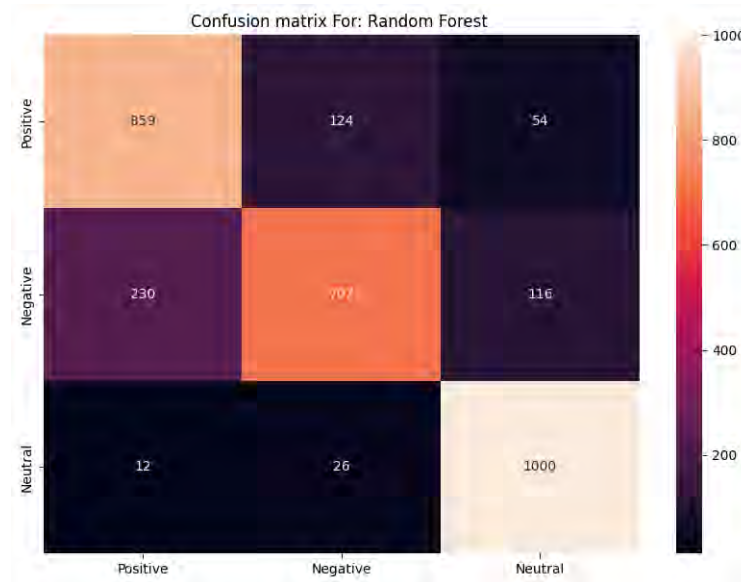


Figure 4.2: Confusion Matrix for Basic Random Forest Model

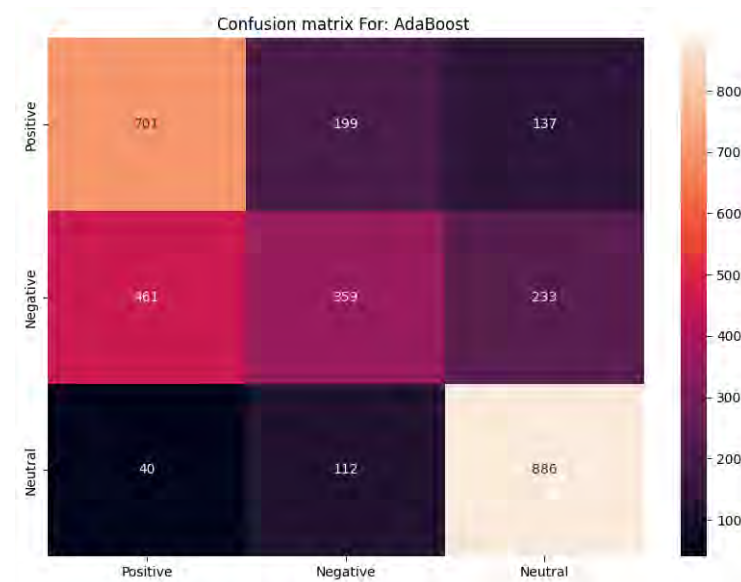


Figure 4.3: Confusion Matrix for Basic AdaBoost Model

### Models using the Weighted Feature Importance Technique

Observing the detailed performance of the models with weighted feature importance, Random Forest models achieve an accuracy of 72% with the weighted feature importance technique. The highlighted fact regarding this is, this performance is achieved

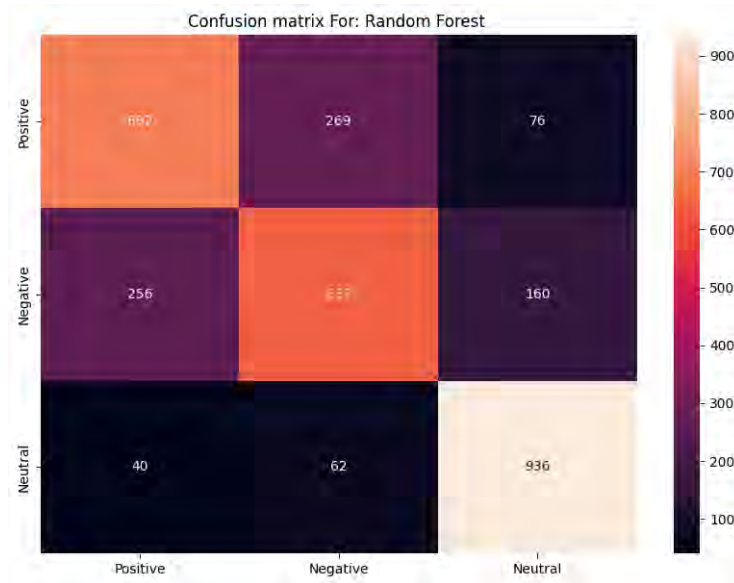


Figure 4.4: Confusion Matrix for Weighted Random Forest Model



Figure 4.5: Confusion Matrix for Weighted AdaBoost Model



with just 15 features. The precision, recall, and F1 scores are also 71%, 72%, and 71% respectively which are pretty good considering the fact that the model was trained with unlabeled data and only 15 features. Coming to the AdaBoost model, it is seen that the model achieves 62% accuracy on the test set with 15 features. The weighted precision of the AdaBoost model is 61%. From the confusion matrices, we can see that the AdaBoost model identifies the neutral class most well by 775 data points but misclassifies most of the Positive and Negative data. The main issue lies on the confusion between positive and negative classification for this model. While the random forest model classifies the neutral and positive features accurately with 936 and 692 data respectively, it fails to classify nearly 400 negative features.

### **Overall Result Comparison on Test Set**

The table 4.2 shows the results for the proposed system in comparison to the basic model. Overall, it is seen that the traditional Random Forest performs best with the test dataset but that is quite normal as they learn from the labeled dataset with whole features. On the other side, if we look at the performance of our proposed system, only implementing the SHAP feature importances itself outstands the performance of traditional AdaBoost with just 15 features compared to the 50 features of the traditional model. The system also performs nearly the same with the combined feature selection of SHAP and estimators and with the weighted feature selection technique.

Table 4.2: Overall Result Comparison of Unimodal Models on Test Set

Method	Model	Feature No	Training Data Type	Accuracy (%)	Precision (%)	Recall (%)	F-1 Score (%)
Traditional	RF	50	100% Labeled	82	82	82	82
	AdaBoost	50	100% Labeled	62	61	62	60
SHAP Feature Selection	RF	15	20% Labeled 80% Unlabeled	70	70	71	70
	AdaBoost	15	20% Labeled 80% Unlabeled	<b>66</b>	65	66	65
Combined Feature Selection	RF	15	20% Labeled 80% Unlabeled	<b>72</b>	72	72	72
	AdaBoost	15	20% Labeled 80% Unlabeled	61	60	61	61
Weighted Feature Selection	RF	15	20% Labeled 80% Unlabeled	<b>72</b>	71	72	71
	AdaBoost	15	20% Labeled 80% Unlabeled	62	61	62	61

The proposed Random Forest model also performs well when trained with 20% labeled and 80% unlabeled data and selecting only 15 features with the weighted technique. In all three feature selection methods it achieves more than 70% accuracy. When the weighted feature selection technique is employed the Random forest model achieves the highest 72% of accuracy with the same percentages of F1 Score. From the detailed results, it can be said that the models perform very well with a very less amount of labeled data and are able to learn from these and predict the unlabeled data. The feature importance also plays an important role in this where with only 15 features the models achieve nearly the same or even higher accuracy than the models who are trained with all of the features. As the Random Forest Performed best with the Unimodal systems we have used the Random Forest as the base model for our Proposed multimodal model.

## 4.4 Results of Multimodal Systems

For multimodal systems, three modalities have been considered - Audio, Text, and Image. The different modalities have been experimented with in combination with audio speech modality in this study. The detailed experimental results are given below -

### 4.4.1 Audio Text Modal

#### Random Forest

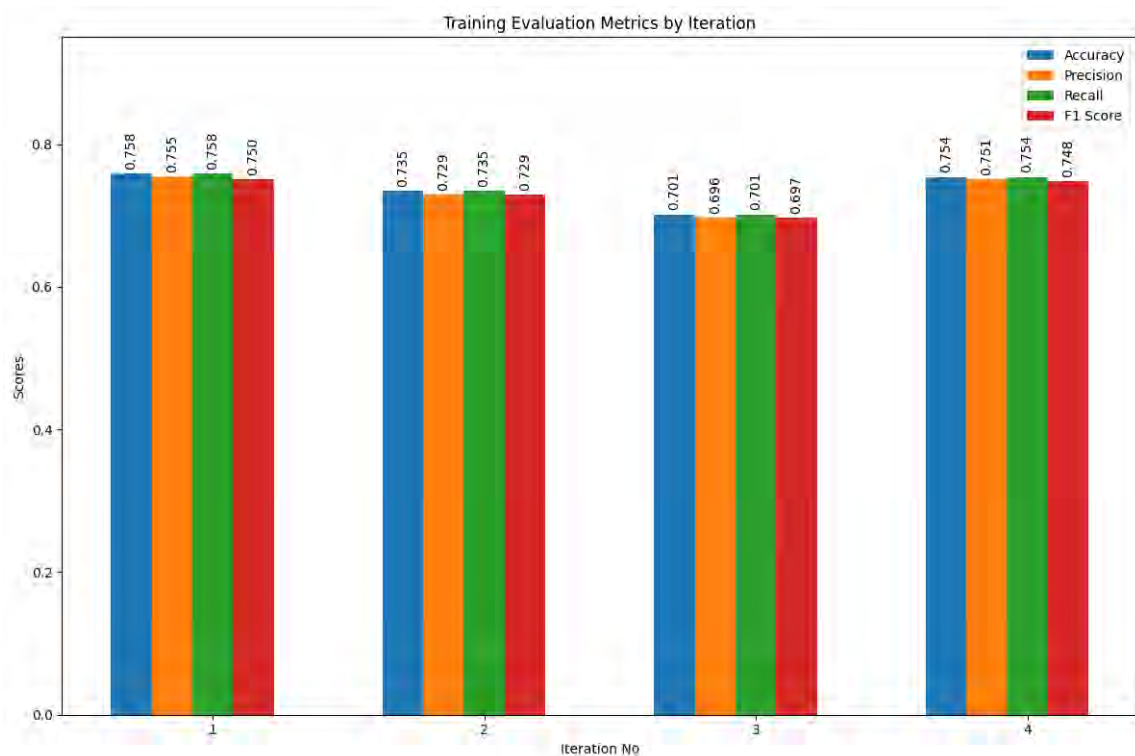


Figure 4.6: Random Forest Training Evaluation Metrics for Audio Text Modal

For Audio and Text modality, the Enhanced Random Forest model used with an

Table 4.3: Classification Report of Proposed ML Model for Multimodal(Audio and Text) Approach

	<b>Precision</b>	<b>Recall</b>	<b>F-1 Score</b>
<b>Positive</b>	0.68	0.71	0.70
<b>Negative</b>	0.72	0.57	0.63
<b>Neutral</b>	0.76	0.90	0.83
<b>Accuracy</b>	0.72		
<b>Macro Average</b>	0.72	0.73	0.72
<b>Weighted Average</b>	0.72	0.72	0.72

unimodal system has also been experimented with. We can see the training evaluation metrics during iterations from the figure 4.6. In the first iteration, the model gets an accuracy of 75.8% and it decreases to 70% for the next two iterations. Finally, at the 5th iteration, the model achieves a training accuracy of 75.4%. In the test results, we can observe from the confusion matrix figure 4.7 is like the unimodal system the model performs well with positive and neutral sentiments but misclassifies 286 negative sentiment data as positive which affects the model's performance. The model performs best with neutral sentiment which is shown by 906 correctly classified data points. 296 positive sentiments are wrongly classified as others. From the classification report Table 4.3, it is seen that the individual negative sentiment has an F1 score of 63% which is less than the other classes. The neutral class has the highest F1 score of 83%.

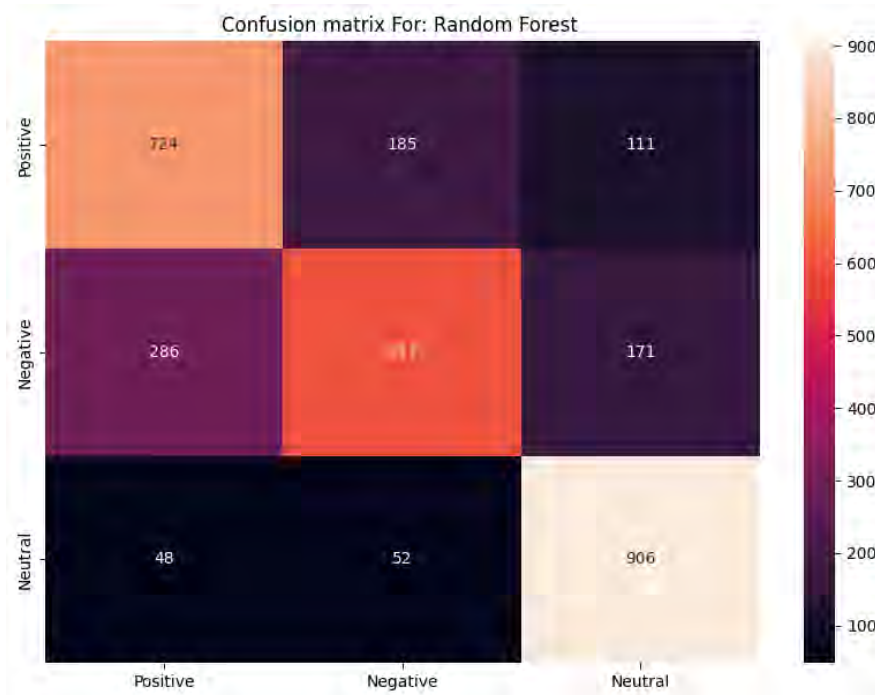


Figure 4.7: Confusion Matrix of Random Forest for Audio Text Modal

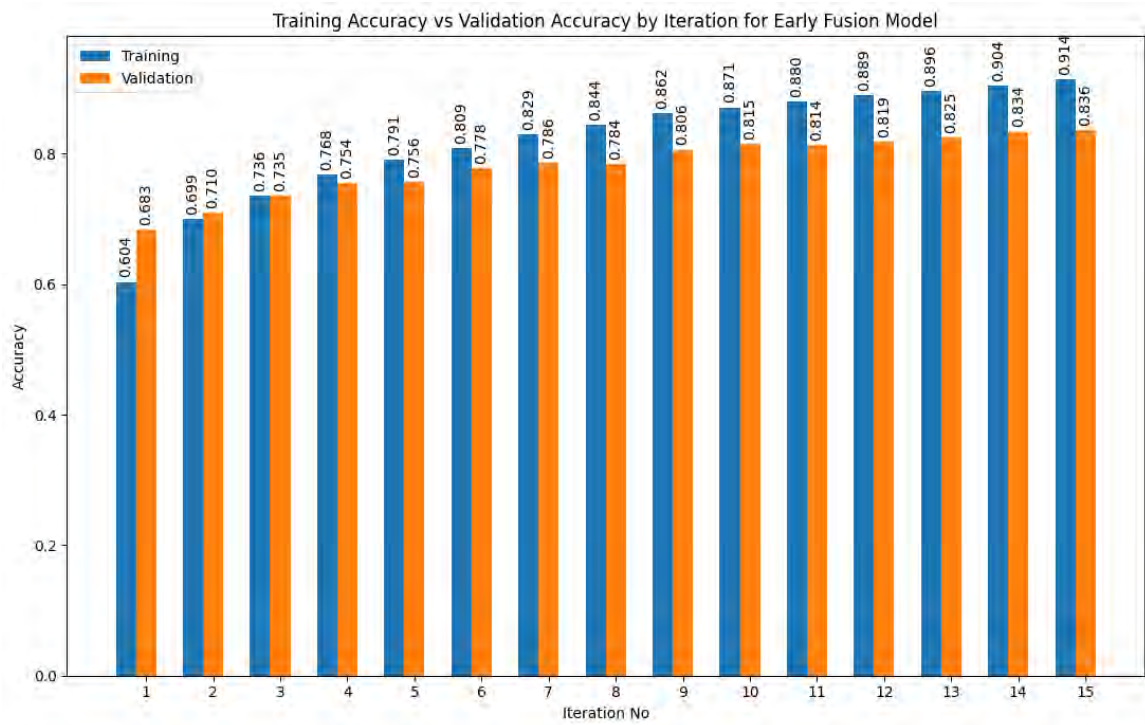


Figure 4.8: Early Fusion Training vs Validation Accuracy for Audio Text Modal

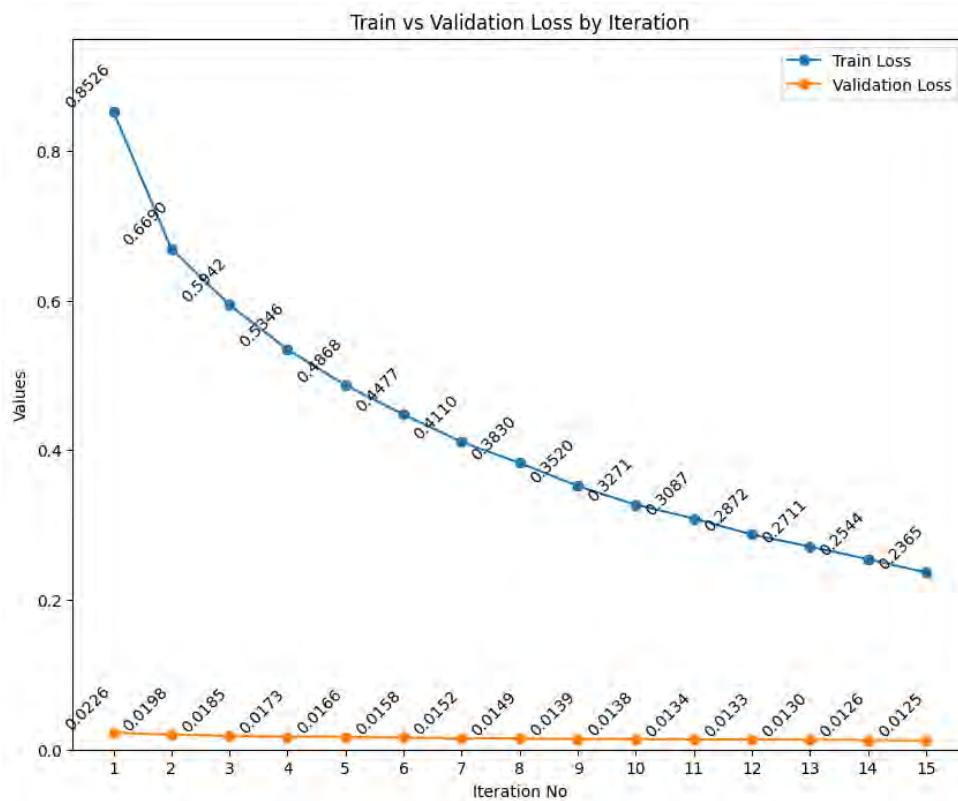


Figure 4.9: Early Fusion Training and Validation Loss for Audio Text Modal

Table 4.4: Classification Report of Early Fusion Deep Learning Model for Multimodal(Audio and Text) Approach

	Precision	Recall	F-1 Score
<b>Positive</b>	0.79	0.84	0.81
<b>Negative</b>	0.81	0.70	0.75
<b>Neutral</b>	0.88	0.94	0.91
<b>Accuracy</b>	0.83		
<b>Macro Average</b>	0.83	0.73	0.82
<b>Weighted Average</b>	0.72	0.72	0.82

### Early Fusion

In the Early fusion technique, the training went for 15 epochs where the training accuracy started from 60% which went up to 91% in the 15th epoch as shown in Figure 4.8. The validation accuracy was higher than the training accuracy at the start but gradually decreased. It achieved the highest accuracy of 83% finally. The validation loss was drastically less at the beginning compared to the training loss as shown in Figure 4.9. The training loss also gradually decreased iteration by iteration. In the evaluation of the test set, the model achieved an accuracy of 83%. The neutral class had the highest score of 91% whereas the negative sentiment class also had a better score of 75% compared to the Random forest model. From the confusion matrix 4.10, we can see that a total of 856 positive sentiments are correctly classified, and the negative and neutral classes had 742 and 947 corrects respectively.



Figure 4.10: Confusion Matrix of Early Fusion Technique for Audio Text Modal

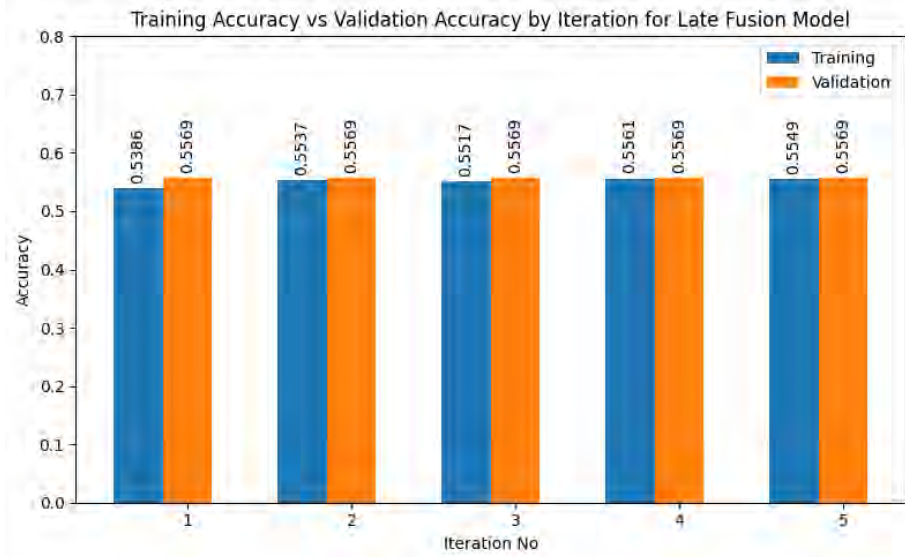


Figure 4.11: Late Fusion Technique Training vs Validation Accuracy Evaluation Metrics for Audio Text Modal

Table 4.5: Classification Report of Late Fusion Model for Multimodal(Audio and Text) Approach

	<b>Precision</b>	<b>Recall</b>	<b>F-1 Score</b>
<b>Positive</b>	0.47	0.62	0.53
<b>Negative</b>	0.74	0.40	0.52
<b>Neutral</b>	0.37	0.80	0.51
<b>Accuracy</b>	0.52		
<b>Macro Average</b>	0.53	0.60	0.52
<b>Weighted Average</b>	0.61	0.52	0.52

## Late Fusion

The late fusion model performance for the Audio text model is not that satisfactory compared to the early fusion and Random forest model. It achieved an accuracy of 52% for the test set. From the training phase as shown in Figure 4.11, it had low accuracy for both the training and validation sets and it didn't increase the accuracy during iterations. From the confusion matrix 4.12, it is clear that the model failed to identify most of the negative sentiments and misclassified them. It correctly classified the positive sentiment for 335 data only.

### 4.4.2 Audio Image Modal

#### Random Forest

For Audio Image modality in figure 4.13, the random forest model was trained for seven iterations until the stopping criteria were satisfied. The accuracy, precision, recall, and f1 score reached 72% at the final iteration. During the training, the second iteration had the highest accuracy of 73%. During the testing phase, the model performance was the same as the training phase with an accuracy of 73% in total. The classification report in table 4.9 shows that the model has the best



Figure 4.12: Confusion Matrix of Late Fusion Technique for Audio Text Modal

Table 4.6: Classification Report of Proposed ML Model for Multimodal(Audio and Image) Approach

	Precision	Recall	F-1 Score
<b>Positive</b>	0.69	0.73	0.71
<b>Negative</b>	0.70	0.53	0.61
<b>Neutral</b>	0.78	0.92	0.84
<b>Accuracy</b>	0.73		
<b>Macro Average</b>	0.72	0.73	0.72
<b>Weighted Average</b>	0.61	0.73	0.72

performance for neutral sentiment classification with an 84% f1 score where 906 data are correctly classified. On the other hand, the model classified 594 negative data correctly resulting in a low 61% F1 score.

### Early Fusion

Like Audio Text modality, the performance of the early fusion model with Audio Image modality is also satisfactory. During the training phase, the model had higher training accuracy which can be seen in the figure 4.22. The training accuracy was 0.86 when the model started training and reached 0.907 when the training stopped with a validation accuracy of 0.815. The validation accuracy also increased over iterations. The validation loss was 0.01 at the final iteration compared to the 0.2611 training loss in figure 4.23. The performance of the model was nearly the same during the testing when the model's final accuracy was 81%. The model correctly classifies a total of 988 Neutral speech data and 833 positive speech data while the model misclassified a total of 342 negative speech data as shown in the confusion matrix in Figure 4.17. Most of the negative speech data were misclassified as positive. The



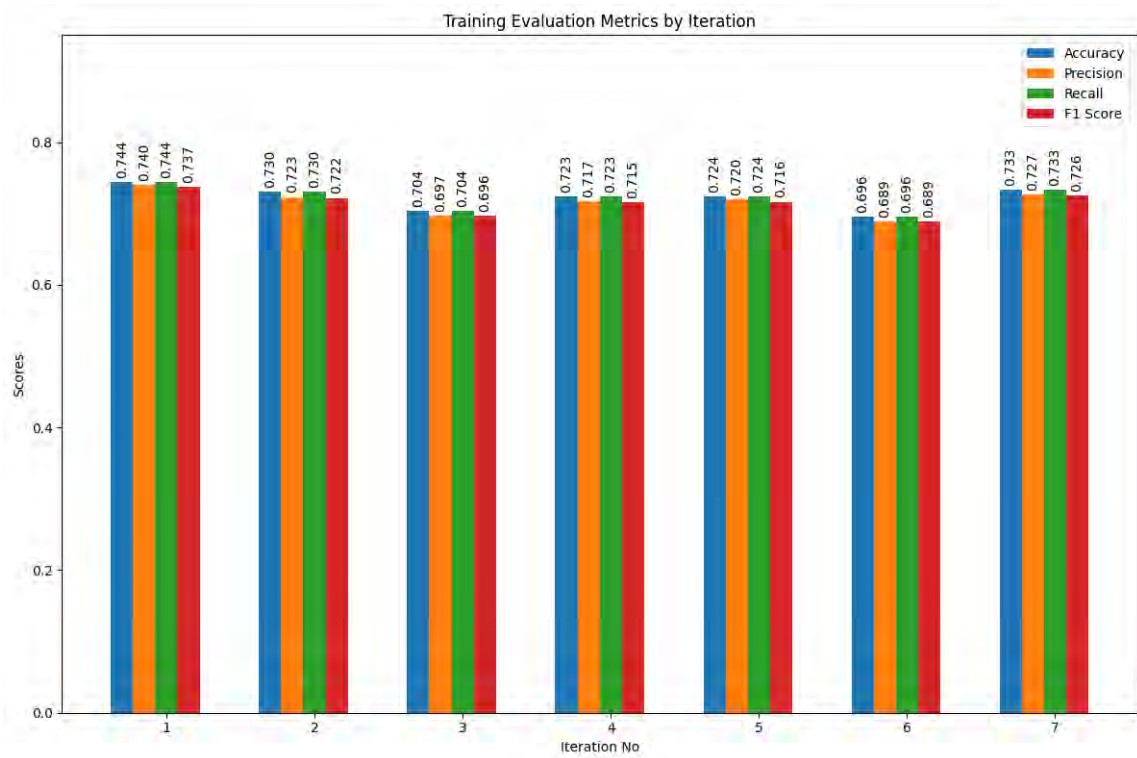


Figure 4.13: Random Forest Training Evaluation Metrics for Audio Image Modal

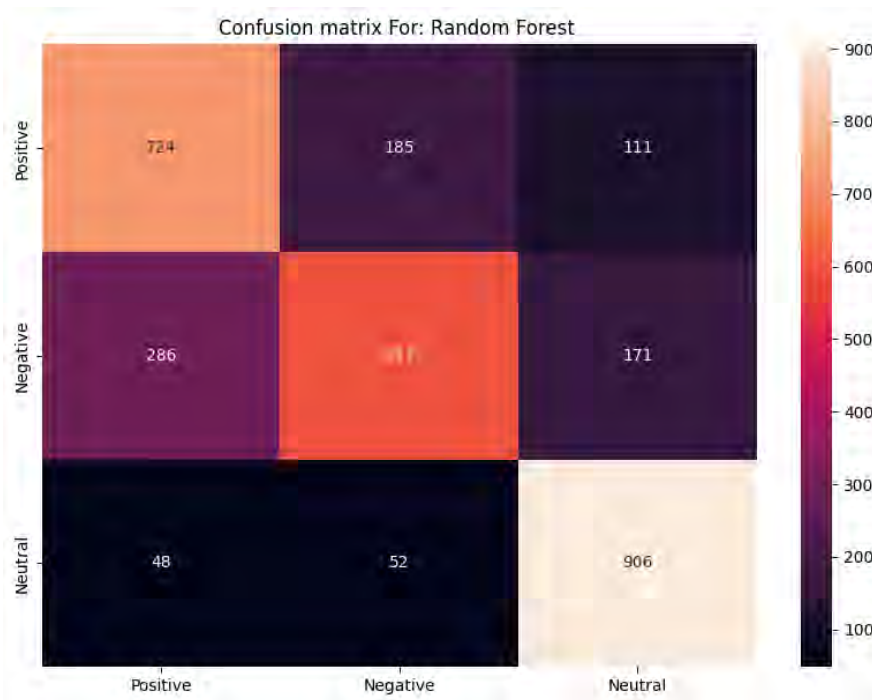


Figure 4.14: Confusion Matrix of Random Forest for Audio Image Modal

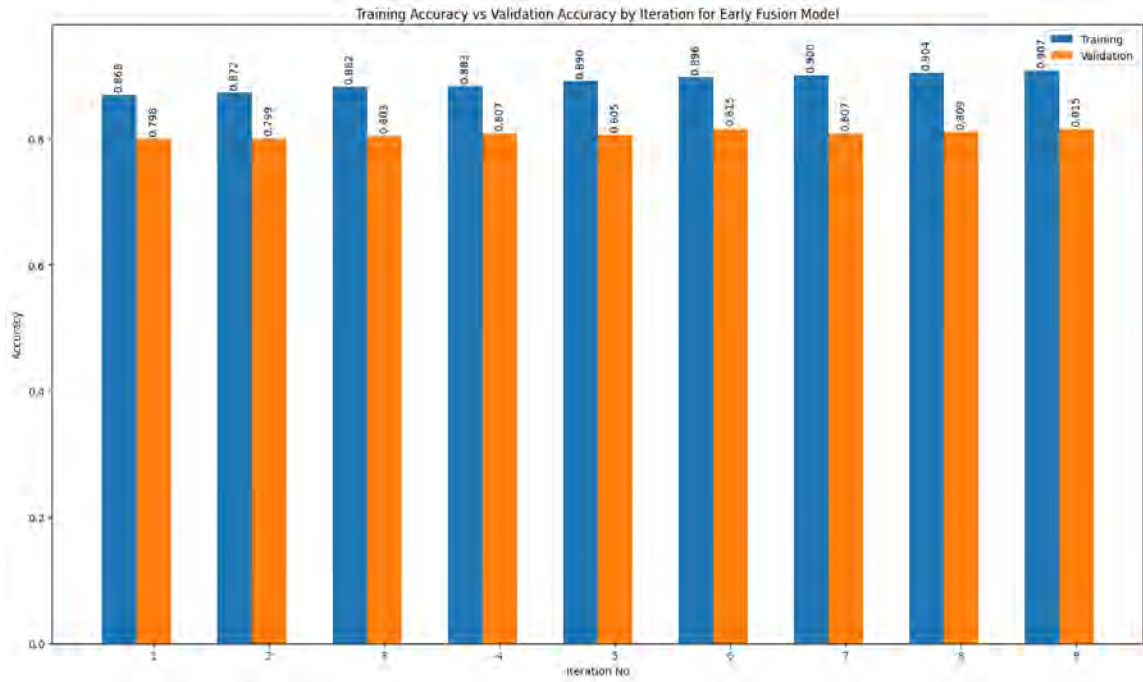


Figure 4.15: Early Fusion Training vs Validation Accuracy for Audio Image Modal

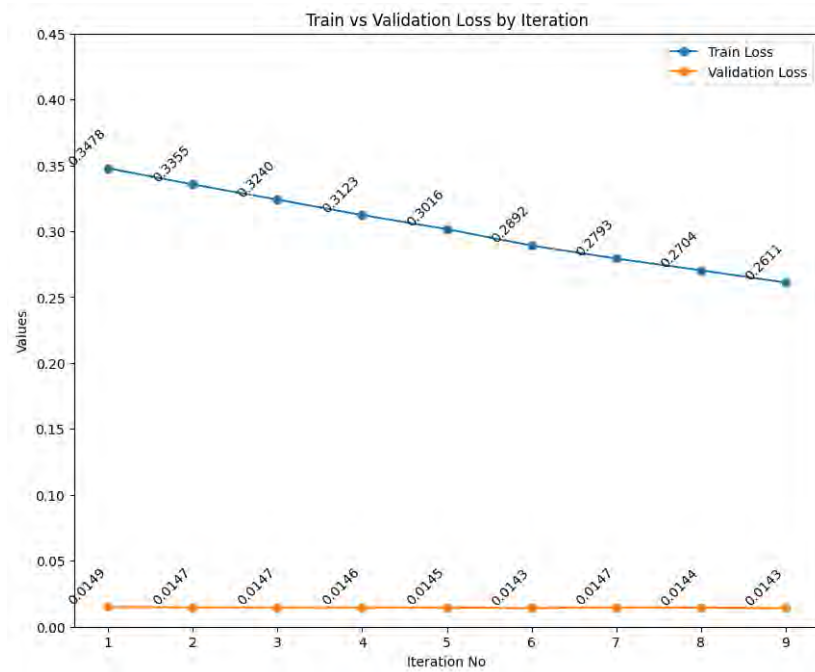


Figure 4.16: Early Fusion Training and Validation Loss for Audio Image Modal

Table 4.7: Classification Report of Early Fusion Deep Learning Model for Multi-modal(Audio and Image) Approach

	Precision	Recall	F-1 Score
<b>Positive</b>	0.76	0.80	0.78
<b>Negative</b>	0.80	0.68	0.73
<b>Neutral</b>	0.86	0.95	0.91
<b>Accuracy</b>	0.81		
<b>Macro Average</b>	0.81	0.81	0.81
<b>Weighted Average</b>	0.81	0.81	0.81

classification report in the figure also shows that a 91% f1 score was recorded for the neutral class and 78% and 73% for the positive and negative classes respectively.

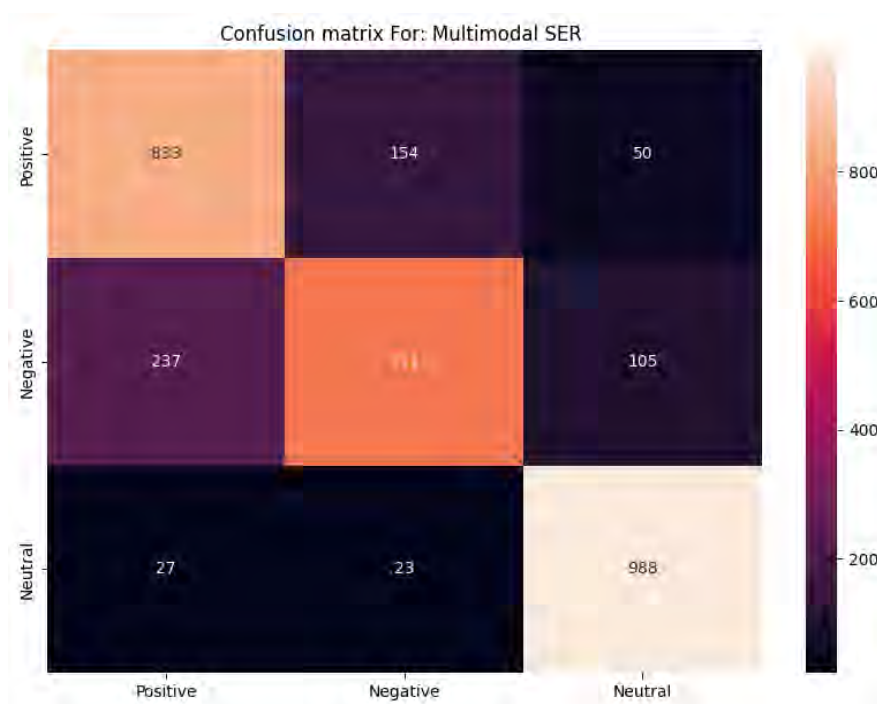


Figure 4.17: Confusion Matrix of Early Fusion Technique for Audio Image Modal

### Late Fusion

In the late fusion model we can see from figure 4.18, the performance increased compared to that of Audio text modality. When audio image features are separately fed to LSTM and CNN models, on the first iteration the training and validation accuracy was poor at 40% and 48% respectively. However, as we can see from figure 4.18 both of the accuracies increased over iterations and reached 65% of validation accuracy when the validation loss was not decreasing any more. At the test phase, the model identified nearly 647 negative speech samples correctly which made a 68% F1 score for the negative class as seen in the table 4.8. Also, the model classified 169 positive sentiment speech data as negative which reduced the F1 score for the class by 58%.

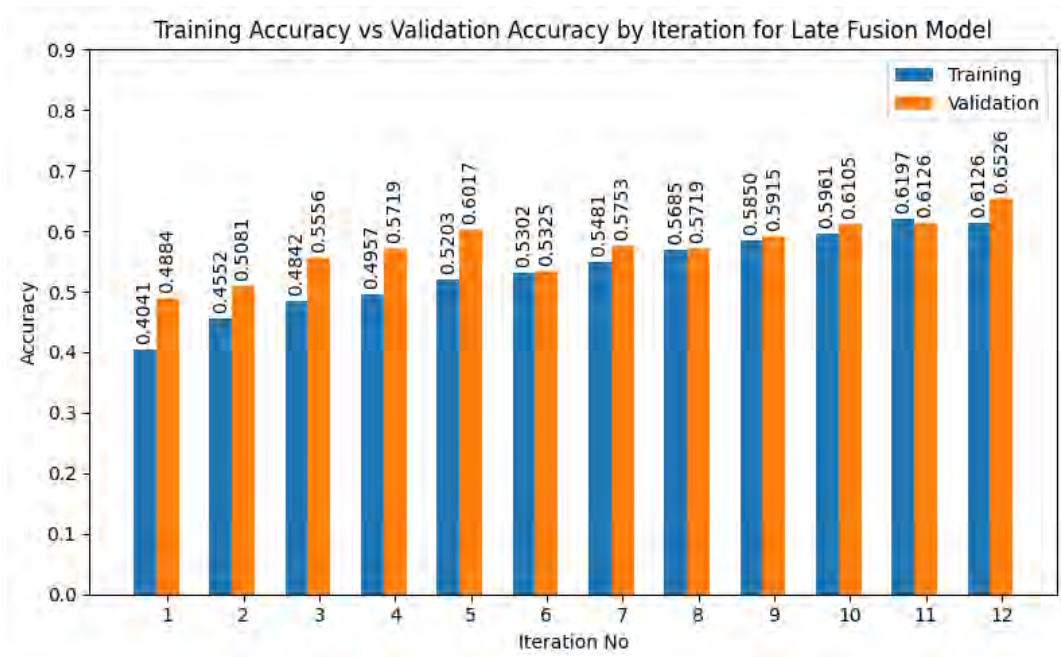


Figure 4.18: Late Fusion Technique Training vs Validation Accuracy Evaluation Metrics for Audio Image Modal



Figure 4.19: Confusion Matrix of Late Fusion Technique for Audio Image Modal

Table 4.8: Classification Report of Late Fusion Model for Multimodal(Audio and Image) Approach

	Precision	Recall	F-1 Score
<b>Positive</b>	0.54	0.62	0.58
<b>Negative</b>	0.74	0.63	0.68
<b>Neutral</b>	0.51	0.67	0.58
<b>Accuracy</b>	0.63		
<b>Macro Average</b>	0.60	0.81	0.61
<b>Weighted Average</b>	0.65	0.81	0.63

### 4.4.3 Audio Text Image Modal

#### Random Forest

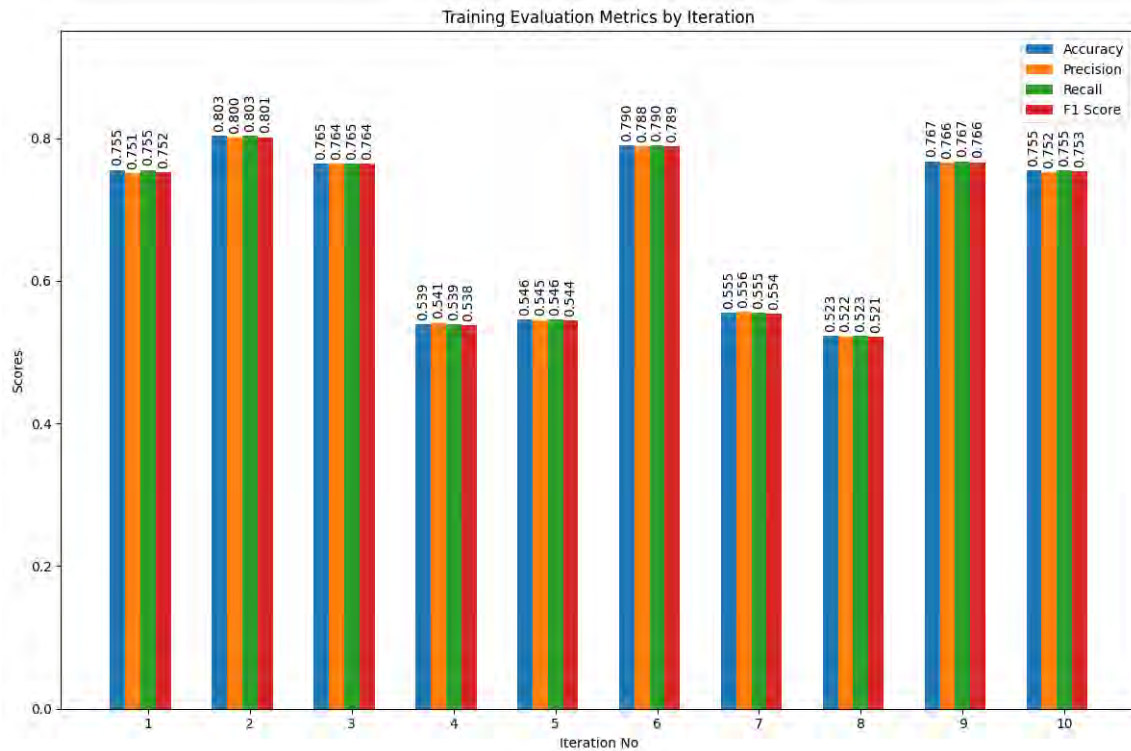


Figure 4.20: Random Forest Training Evaluation Metrics for Audio Text and Image Modal

When all three modalities are used in our proposed system, the Random Forest model performed relatively well. From figure 4.20 we can see, during the training time, the F1 score was nearly 75% when the iterations started and it increased to 80% on the next iteration. However, the trend started decreasing for a while and finally increased at the 6th iteration. But the performance was not stable and had frequent changes when at the last step before stopping it reached 75%. When we tested the model with our test set the model achieved an accuracy of 77% with nearly 91% of individual F1 score for the neutral class. For the negative class, the model had an F1 score of 68%. From the confusion matrix in figure 4.21, we can see that the model predicted a total of 597 speech sentiments as negative sentiments correctly

while misclassifying others. A good amount of Neutral (906) and positive(724) data are also correctly classified. We can see that the wrong classification of Positive and negative sentiments is affecting the performance of the overall model.

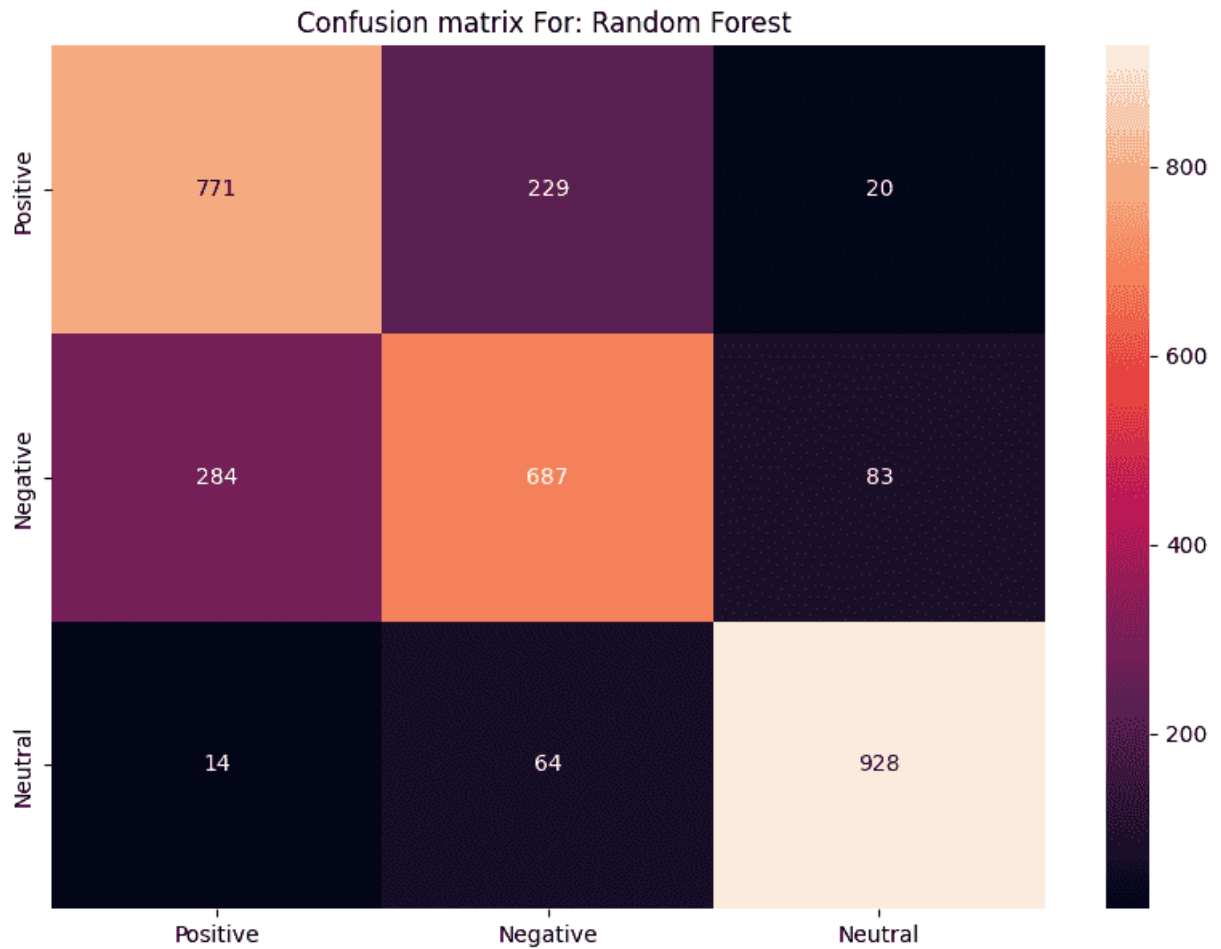


Figure 4.21: Confusion Matrix of Random Forest for Audio Text and Image Modal

### Early Fusion

In the early fusion model when we combined all the features from audio, text, and image together and sent it an LSTM model, we can see in figure 4.22 the training phase starts with a training accuracy of 0.601 but the validation accuracy is slightly higher at 0.64. Gradually, the model learns the features and boosts its performance. The training accuracy shows an increasing trend throughout the training phase but the validation accuracy is not increased much compared to that. Also, as the accuracy gradually increases the training loss also decreases as shown in the figure. Compared to the training loss, the validation loss remains nearly stable with small amount of decrease and stays in a range of 0.0236 to 0.016 as shown in figure 4.23. The initial behavior reverts when from the 4th iteration we can see that the model’s validation performance(0.73) becomes less than the training performance(0.75). At the end of training when the model stops training, the final accuracy for training is nearly 91% and 79% as validation accuracy. When we evaluated the trained model with our test set we received a final weighted accuracy of 79% and other metrics had

Table 4.9: Classification Report of Proposed ML Model for Multimodal(Audio, Text and Image) Approach

	Precision	Recall	F-1 Score
<b>Positive</b>	0.72	0.76	0.74
<b>Negative</b>	0.70	0.65	0.68
<b>Neutral</b>	0.90	0.92	0.91
<b>Accuracy</b>	0.77		
<b>Macro Average</b>	0.77	0.78	0.77
<b>Weighted Average</b>	0.77	0.77	0.77

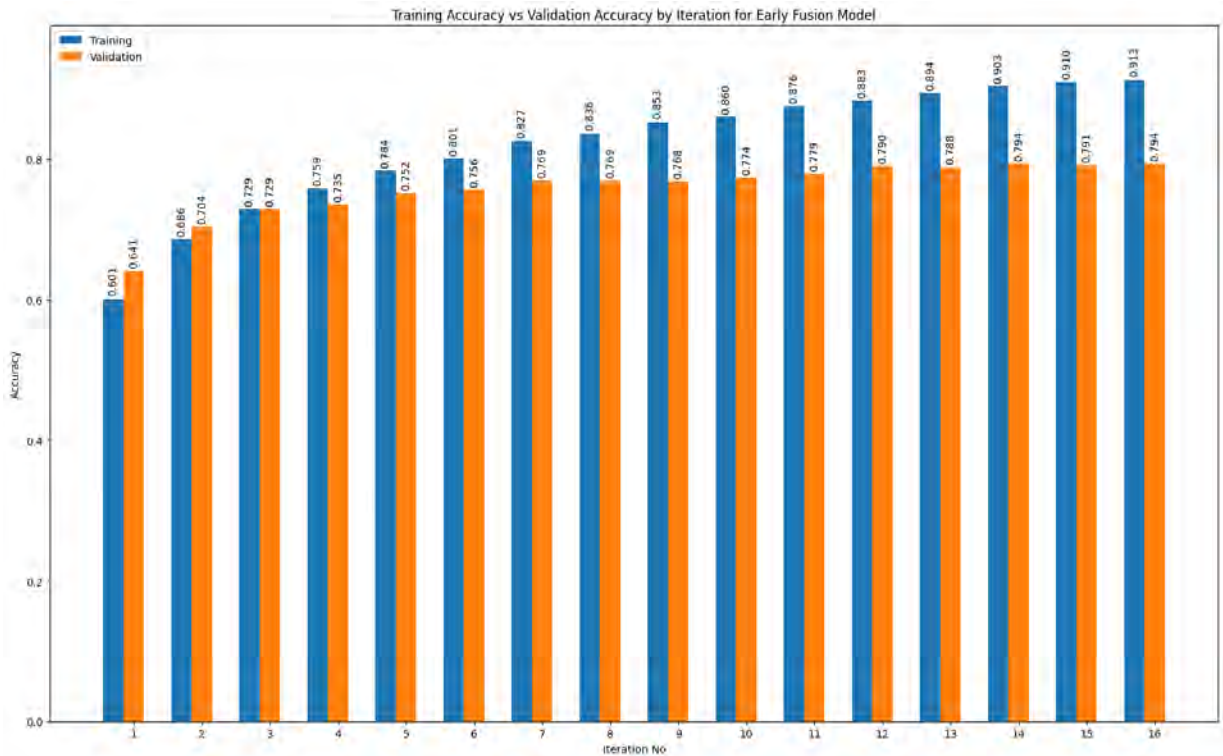


Figure 4.22: Early Fusion Training vs Validation Accuracy for Audio, Text and Image Modal

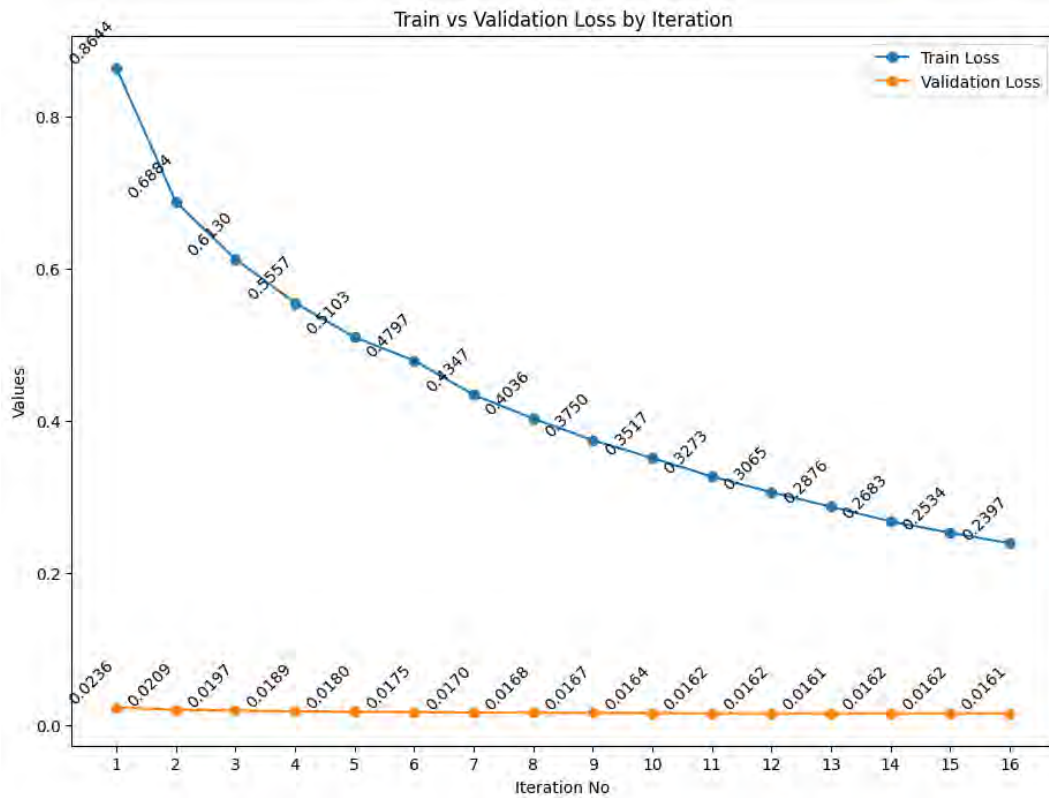


Figure 4.23: Early Fusion Training and Validation Loss for Audio, Text and Image Modal

the same result. The recall of the Neutral class is the highest which is 0.93 among the three classes as shown in table 4.10. The model was able to correctly classify 905 neutral classes. Moreover, 196 negative speeches were predicted as positive, and 121 negative speeches were predicted as neutral which resulted in to a 0.69 recall score for that particular class.

### Late Fusion

When training the model with separate LSTM, CNN, and BanglaBERT classification models and combining the result to go through a fully connected layer, we can see from figure 4.25 that the training and validation accuracy is not quite well compared to the early fusion model. The training and validation accuracy for the first

Table 4.10: Classification Report of Early Fusion Deep Learning Model for Multi-modal(Audio, Text and Image) Approach

	Precision	Recall	F-1 Score
<b>Positive</b>	0.79	0.75	0.77
<b>Negative</b>	0.74	0.69	0.72
<b>Neutral</b>	0.83	0.93	0.88
<b>Accuracy</b>	0.79		
<b>Macro Average</b>	0.79	0.79	0.79
<b>Weighted Average</b>	0.79	0.79	0.79





Figure 4.24: Confusion Matrix of Early Fusion Technique for Audio, Text and Image Modal

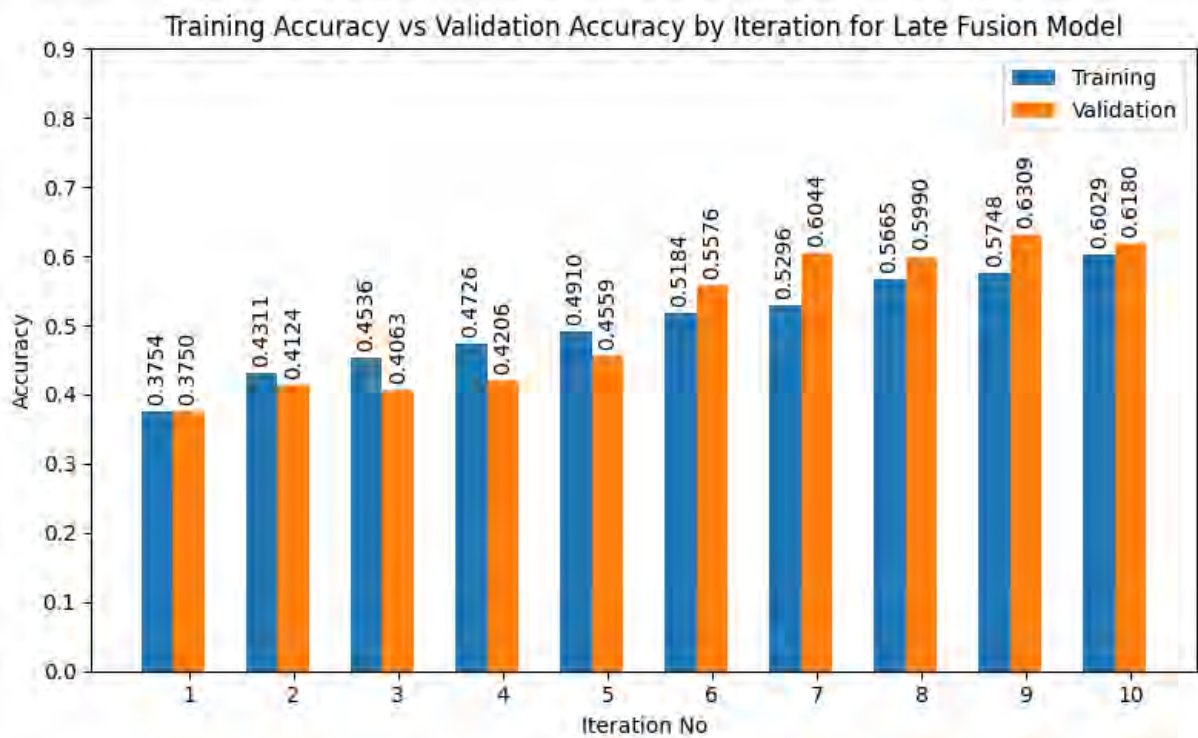


Figure 4.25: Late Fusion Technique Training vs Validation Accuracy Evaluation Metrics for Audio Text and Image Modal

Table 4.11: Classification Report of Late Fusion Model for Multimodal(Audio, Text and Image) Approach

	Precision	Recall	F-1 Score
<b>Positive</b>	0.54	0.62	0.58
<b>Negative</b>	0.74	0.63	0.68
<b>Neutral</b>	0.51	0.67	0.58
<b>Accuracy</b>	0.63		
<b>Macro Average</b>	0.60	0.64	0.61
<b>Weighted Average</b>	0.65	0.63	0.63

3 iterations were very low which started with only 0.375. Until the 5th iteration, the validation performance was less than the training set performance. From the 6th iteration, the behavior reversed. Nonetheless, we can see an increasing trend throughout the training phase for the mode where the training was stopped after the 10th iteration as the validation loss was not increasing as shown in the figure. The final training and validation accuracy of the model was 0.602 and 0.618 respectively.

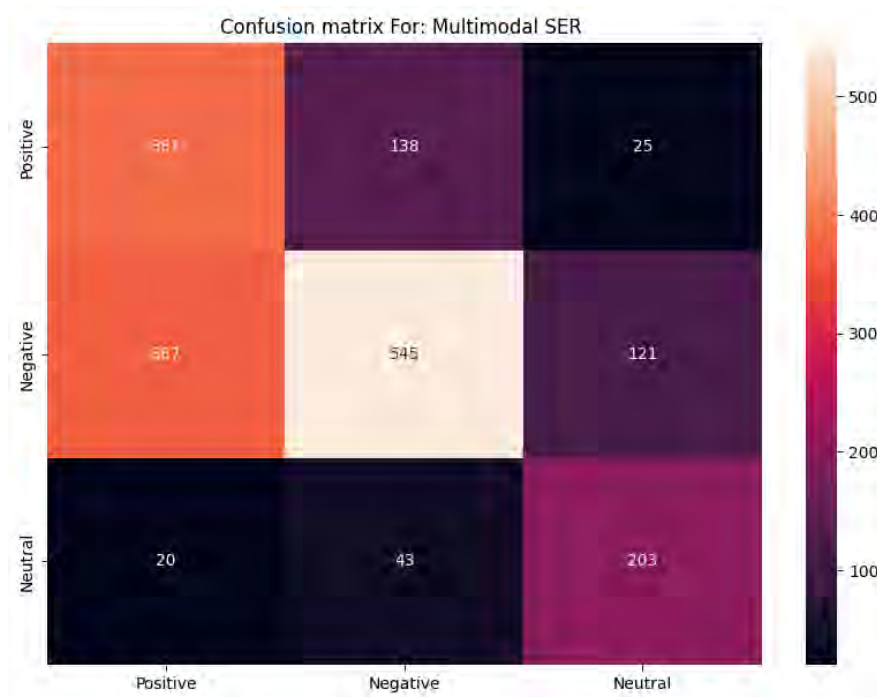


Figure 4.26: Confusion Matrix of Late Fusion Technique for Audio, Text and Image Modal

For the testing performance, from table 4.11 it is seen that the model achieved nearly the same performance as the validation set which is 63% of accuracy, and the weighted precision of the model was highest with a 0.63 score. The F1 score for individual positive and neutral classes was 0.58 and the supported data is also less for these two classes. From the confusion matrix in figure 4.26, we can see that 381 speech sentiments were correctly classified as positive and 545 data were accurately classified as negative. The neutral class had the least misclassifications which is only 63 in number.

#### 4.4.4 Combined Result Analysis on Test Set

From the table 4.12, we can see a detailed performance comparison of all the multimodal speech sentiment recognition systems that have been experimented on. It is evident that the early fusion technique worked well for the sentiment recognition algorithms achieving nearly 80% or higher performance in terms of all the evaluation metrics. Overall, the sentiment recognition works best with Speech and text modalities employing LSTM as the classifier and achieves the highest 83% accuracy on the test set. In contrast for the late fusion technique, speech, and text modalities seem to underperform with only 54% accuracy and precision is also less than 53%. Compared to that the Speech and Image modality performs better than all other late fusion models with 62% of accuracy. Moreover, when all three modalities (Speech, Text, Image) are used the model has relatively stable performance compared to all other models. This shows the importance of using multimodal systems where each modality compresses the downside of other modalities and provides a generalized performance. The achieved performance of this experiment was 77%, 79%, and 61% respectively for the early fusion technique using Random Forest, LSTM, and late fusion technique using LSTM for Audio, CNN for image, and BanglaBert for textual features.

Table 4.12: Combined Result Analysis on Test Set for Multimodal System

Multimodality	Technique	Model	Feature No	Learning Type	Accuracy (%)	Precision (%)	Recall (%)	F-1 Score (%)
Text & Speech	Early Fusion	SHAP+RF	20	Semi-Supervised	72	72	72	72
		Word2Vec+LSTM	130	Supervised	<b>83</b>	83	83	81
Image & Speech	Late Fusion	Text- Bangla BERT	Audio-20	Supervised	54	53	54	54
		Audio- LSTM	Text-18					
	Early Fusion	SHAP+RF	20	Semi-Supervised	73	72	73	72
		VGG19+LSTM	51	Supervised	<b>81</b>	81	81	81
Late Fusion	Image- CNN	Audio-20		Supervised	62	61	62	62
	Audio-LSTM							
Speech, Text & Image	Early Fusion	SHAP+RF	20	Semi-Supervised	77	77	77	77
		VGG19+						
	Word2Vec+ LSTM	Audio, Text-151	Supervised	<b>79</b>	79	79	79	
Late Fusion	Text - Bangla BERT	Audio, Text-151		Supervised	61	65	61	61
	Image - CNN							
		Audio - LSTM	Audio, Text-151	Supervised				

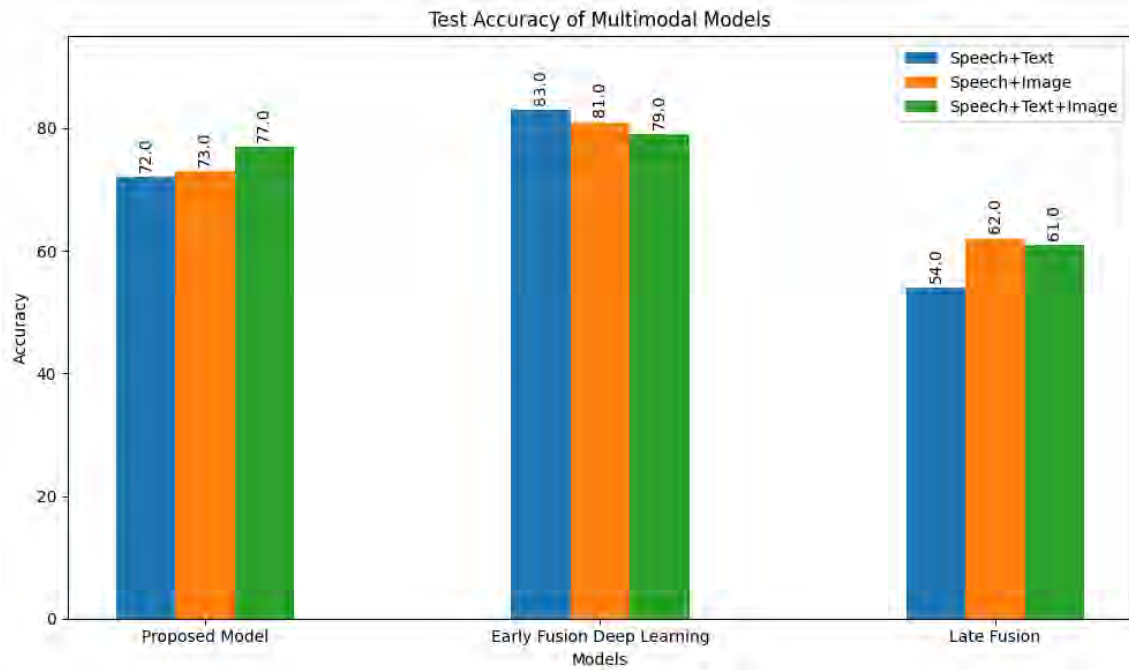


Figure 4.27: Test Accuracy Comparison Among Different Models

#### 4.4.5 Comparison of Random Forest and LSTM Model for Semi-Supervised Learning

From the table 4.12 it is visible that the semi-supervised model performed best with Audio, Image, and Text modality using the Early Fusion Technique. So we experimented with the same modality and fusion technique but with a different base model for the semi-supervised learning. As LSTM performed best in the multimodal model so we used LSTM for the comparison. In table 4.13 we can see the performance comparison. It is seen that, when we use the LSTM as the base model in the semi-supervised loop, the model is trained for 23 epochs whereas Random Forest was trained for 16 epochs. The LSTM model achieved an accuracy of 69% on the test set which is lower than the 77% accuracy of the Random Forest model. In terms of execution time also, Random Forest model took less time compared to LSTM.

Table 4.13: Performance Comparison between RF and LSTM for Semi-Supervised Multimodal Model (Audio, Image and Text)

Model	Epochs	Accuracy (%)	Precision (%)	Recall (%)	F-1 Score (%)	Execution Time
Random Forest	16	77	77	77	77	13m 18s
LSTM	23	69	69	69	69	37m 27s

## 4.5 Discussion

We can see the final performance comparison of Multimodal models in Figure 4.27. One notable thing to mention in the overall performance is the satisfactory performance of the proposed enhanced multimodal Random Forest model. From the table, we can see that the Random Forest model showed consistent performance for all three combinations of modalities (Speech Text, Speech Image, and Speech, text, and image). It achieved an accuracy of 72%, 73%, and 77% respectively. Another impacting fact is that the model achieved this performance with only 20 features combined from different modalities compared to the results of other models trained with 100+ features. Moreover, Random Forest not only showed satisfactory performance with fewer features it learned the features from unlabeled data. This shows the impact of employing an enhanced semi-supervised learning loop with SHAP feature importance. It can be said that when the proposed enhanced multimodal model is employed the system predicts the sentiments correctly with less dimensionality and is effective on unlabeled data. The performance of the traditional Machine learning model outperformed the deep learning model in terms of multimodality in some cases. Especially for the late fusion techniques, the overall performance is less than early fusion techniques where the highest accuracy achieved for late fusion technique is 62%, while early fusion techniques display a performance of more than 70% in general. In the early fusion technique also, the performance comparison shows the better performance of the Machine learning model compared to the deep learning model in terms of accuracy, precision, recall, F-1 Score, and even the execution time. The machine learning models being lightweight and simple in structure can perform better for Bengali sentiment recognition when enhanced with correct parameters in comparison to the data-dependant, time and resource costly deep learning models.

# Chapter 5

## Conclusion

This study delves deeply into the complexity of Bangla speech sentiment recognition where enhanced machine learning models through SHAP-based semi-supervised learning and both unimodal and multimodal deep learning strategies are employed. A total of seventeen approaches including eight for unimodal systems and nine for multimodal systems are presented in this study. The importance of this research lies in its ability to transform our understanding and interaction with Bangla voice data in practical applications.

In this study the experiments are done in broadly two parts - Unimodal Systems and Multimodal Systems. This study proposed a novel algorithm for the semi-supervised learning of Machine learning models, specifically Random Forest and AdaBoost where the weighted feature importances of SHAP are taken into account for iterative feature selection. We have experimented with feature selection methods in multiple steps to reduce the dimensionality of the models and effective learning. This proposed method is used for both unimodal and multimodal systems. Specifically for multimodal systems, the study implemented three different modality-dependent models to extract the hidden features of the modalities. Sequential model LSTM for acoustic features, Custom CNN for visual features, and BanglaBERT for Bangla textual features have been used in this study to extract accurate features and fusion them for better performance.

In an unimodal system, the proposed enhanced ML system achieved a satisfactory performance of 71% accuracy with only 20 features. When employing the same model for multimodal systems, this model exhibited consistent performance across various modality combinations (speech-text, speech-image, and speech-text-image), achieving accuracies of 72%, 73%, and 77%, respectively. These findings suggest that our enhanced model is capable of making accurate sentiment predictions with fewer features and is particularly effective with unlabeled data, demonstrating superior performance compared to some deep learning models in multimodal contexts.

Furthermore, our extensive experiments highlight the significance of using multimodal approaches in terms of Bangla Speech sentiment analysis over the unimodal ones. In the experiments, utilizing the early fusion technique for multimodal systems achieved nearly 80% or higher across all evaluation metrics, with the speech and text modalities combined with an LSTM classifier reaching the highest accuracy

of 83%. In contrast, the late fusion techniques demonstrated underperformance, particularly in the speech and text modalities, with accuracy and precision dropping below 54% and 53%, respectively. However, when speech and image modalities were employed in late fusion, a moderate improvement to 62% accuracy was observed. Utilizing all three modalities (speech, text, and image) provided relatively stable performance, highlighting the robustness and generalization capability of multimodal systems where each modality mitigates the weaknesses of the others.

In conclusion, this research provides the foundation for developing robust systems that can analyze the emotional undercurrents of Bangla speech, leading to advancements in various applications. Academically, it paves the way for future studies in low-resource languages by offering a scalable and adaptable framework for sentiment analysis. We believe that this study will pave the way for future studies in low-resource languages by offering a scalable and adaptable framework for sentiment analysis. It will open a new view where the explainable AI can be integrated into the iterative improvement of the traditional ML models instead of leaning towards deep learning models for unlabeled data. Also, it provides insight into using multimodal systems and their effectiveness for specific tasks. By incorporating additional modalities like facial expressions, exploring generalizability across dialects, and refining the SHAP-based approach, we can further enhance the robustness and reach of these techniques. By continuing to develop these methods, we can bridge the gap in understanding the vast and emotionally rich landscape of Bangla speech data being generated today. This will ultimately lead to more effective communication technologies, improved customer service interactions, and deeper insights into the sentiments expressed by Bangla speakers across the globe.

## 5.1 Limitation and Future Work

The most challenging task for this study remains particularly in addressing data scarcity and the nuanced characteristics of the Bangla language. There is an unavailability of the huge amount of data for speech in the Bengali language that can be used for the deep learning models. Future research should focus on expanding datasets, enhancing cross-lingual transfer learning techniques, and exploring real-time applications. Moreover, the proposed algorithms can target different fine-grained emotions under the hood of Positive, Negative, and Neutral sentiments. Additionally, ethical considerations in the development and deployment of AI for sentiment analysis must be continually addressed to ensure responsible use. We plan to build a single robust model by combining all the positive aspects of this study to leverage accurate, efficient, and responsible sentiment analysis.



# Bibliography

- [1] B. P. Yuhas, M. H. Goldstein, and T. J. Sejnowski, "Integration of acoustic and visual speech signals using neural networks," *IEEE Communications Magazine*, vol. 27, no. 11, pp. 65–71, 1989.
- [2] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, pp. 248–255.
- [4] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proceedings of the 13th international conference on multimodal interfaces*, 2011, pp. 169–176.
- [5] N. Altrabsheh, M. M. Gaber, M. Cocea, *et al.*, "Sa-e: Sentiment analysis for education," *Frontiers in Artificial Intelligence and Applications*, vol. 255, pp. 353–362, 2013.
- [6] L. Kaushik, A. Sangwan, and J. H. Hansen, "Sentiment extraction from natural audio streams," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2013, pp. 8485–8489.
- [7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [8] L. Kaushik, A. Sangwan, and J. H. Hansen, "Automatic audio sentiment extraction using keyword spotting.," in *INTERSPEECH*, 2015, pp. 2709–2713.
- [9] D. Bertero, F. B. Siddique, C.-S. Wu, Y. Wan, R. H. Y. Chan, and P. Fung, "Real-time speech emotion and sentiment recognition for interactive dialogue systems," in *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2016, pp. 1042–1047.
- [10] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional mkl based multimodal emotion recognition and sentiment analysis," in *2016 IEEE 16th international conference on data mining (ICDM)*, IEEE, 2016, pp. 439–448.
- [11] H. Abburi, "Audio and text based multimodal sentiment analysis using features extracted from selective regions and deep neural networks," *International institute of information technology Hyderabad-500032, India*, 2017.
- [12] E. Chu and D. Roy, "Audio-visual sentiment analysis for learning emotional arcs in movies," in *2017 IEEE International Conference on Data Mining (ICDM)*, IEEE, 2017, pp. 829–834.

- [13] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in neural information processing systems*, vol. 30, 2017.
- [14] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.
- [15] M. M. Rahman, D. R. Dipta, and M. M. Hasan, “Dynamic time warping assisted svm classifier for bangla speech recognition,” in *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, IEEE, 2018, pp. 1–6.
- [16] S. A. Sumon, J. Chowdhury, S. Debnath, N. Mohammed, and S. Momen, “Bangla short speech commands recognition using convolutional neural networks,” in *2018 international conference on bangla speech and language processing (ICBSLP)*, IEEE, 2018, pp. 1–6.
- [17] M. A. Al Mamun, I. Kadir, A. S. A. Rabby, and A. Al Azmi, “Bangla music genre classification using neural network,” in *2019 8th international conference system modeling and advancement in research trends (SMART)*, IEEE, 2019, pp. 397–403.
- [18] D. M. Eberhard, G. F. Simons, and C. D. Fennig, “Summary by language size,” *SIL International, Ethnologue*, 2019.
- [19] M. Kattel, A. Nepal, A. Shah, and D. Shrestha, “Chroma feature extraction,” in *Conference: chroma feature extraction using fourier transform*, vol. 20, 2019.
- [20] Z. Luo, H. Xu, and F. Chen, “Audio sentiment analysis by heterogeneous signal features learned from utterance-based parallel neural network.,” in *AffCon@AAAI*, Shanghai, China, 2019, pp. 80–87.
- [21] S. Al-Azani and E.-S. M. El-Alfy, “Enhanced video analytics for sentiment analysis based on fusing textual, auditory and visual information,” *IEEE Access*, vol. 8, pp. 136 843–136 857, 2020.
- [22] J. Gao, P. Li, Z. Chen, and J. Zhang, “A survey on deep learning for multimodal data fusion,” *Neural Computation*, vol. 32, no. 5, pp. 829–864, 2020.
- [23] Z. Lu, L. Cao, Y. Zhang, C.-C. Chiu, and J. Fan, “Speech sentiment analysis via pre-trained features from end-to-end asr models,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7149–7153.
- [24] N. Sadeq, N. T. Chowdhury, F. T. Utshaw, S. Ahmed, and M. A. Adnan, “Improving end-to-end bangla speech recognition with semi-supervised training,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 1875–1883.
- [25] B. T. Atmaja and M. Akagi, “Evaluation of error-and correlation-based loss functions for multitask learning dimensional speech emotion recognition,” in *Journal of Physics: Conference Series*, IOP Publishing, vol. 1896, 2021, p. 012 004.
- [26] S. Cunningham, H. Ridley, J. Weinel, and R. Picking, “Supervised machine learning for audio emotion recognition: Enhancing film sound design using audio features, regression models and artificial neural networks,” *Personal and Ubiquitous Computing*, vol. 25, no. 4, pp. 637–650, 2021.

- [27] 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.
- [28] Z. Jiawa, L. Wei, W. Sili, and Y. Heng, “Review of methods and applications of text sentiment analysis,” *Data analysis and knowledge discovery*, vol. 5, no. 6, pp. 1–13, 2021.
- [29] M. R. U. Rashid, M. Mahbub, and M. A. Adnan, “Band: A benchmark dataset for bangla news audio classification,” in *Proceedings of the 3rd ACM International Conference on Multimedia in Asia*, 2021, pp. 1–6.
- [30] M. Sakurai and T. Kosaka, “Emotion recognition combining acoustic and linguistic features based on speech recognition results,” in *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)*, IEEE, 2021, pp. 824–827.
- [31] 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.
- [32] B. Vimal, M. Surya, V. Sridhar, A. Ashok, *et al.*, “Mfcc based audio classification using machine learning,” in *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, IEEE, 2021, pp. 1–4.
- [33] A. Yakaew, M. N. Dailey, and T. Racharak, “Multimodal sentiment analysis on video streams using lightweight deep neural networks,” in *ICPRAM*, 2021, pp. 442–451.
- [34] B. T. Atmaja and A. Sasou, “Sentiment analysis and emotion recognition from speech using universal speech representations,” *Sensors*, vol. 22, no. 17, p. 6369, 2022.
- [35] R. R. Choudhary, G. Meena, and K. K. Mohbey, “Speech emotion based sentiment recognition using deep neural networks,” in *Journal of Physics: Conference Series*, IOP Publishing, vol. 2236, 2022, p. 012 003.
- [36] 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.
- [37] A. Ghriss, B. Yang, V. Rozgic, E. Shriberg, and C. Wang, “Sentiment-aware automatic speech recognition pre-training for enhanced speech emotion recognition,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7347–7351.
- [38] 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.
- [39] K. M. Hasib, A. Tanzim, J. Shin, K. O. Faruk, J. Al Mahmud, and M. F. Mridha, “Bmnet-5: A novel approach of neural network to classify the genre of bengali music based on audio features,” *IEEE Access*, vol. 10, pp. 108 545–108 563, 2022.
- [40] E. Hossain, O. Sharif, and M. M. Hoque, “Memosen: A multimodal dataset for sentiment analysis of memes,” in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 1542–1554.

- [41] E. Hossain, O. Sharif, and M. M. Hoque, “Mute: A multimodal dataset for detecting hateful memes,” in *Proceedings of the 2nd conference of the asia-pacific chapter of the association for computational linguistics and the 12th international joint conference on natural language processing: student research workshop*, 2022, pp. 32–39.
- [42] M. R. Karim, S. K. Dey, T. Islam, M. Shajalal, and B. R. Chakravarthi, “Multimodal hate speech detection from bengali memes and texts,” in *International Conference on Speech and Language Technologies for Low-resource Languages*, Springer, 2022, pp. 293–308.
- [43] R. Shaik and S. Venkatramaphanikumar, “Sentiment analysis with word-based urdu speech recognition,” *Journal of ambient intelligence and humanized computing*, vol. 13, no. 5, pp. 2511–2531, 2022.
- [44] P. Zhao, F. Liu, and X. Zhuang, “Speech sentiment analysis using hierarchical conformer networks,” *Applied Sciences*, vol. 12, no. 16, p. 8076, 2022.
- [45] S. Aziz, N. H. Arif, S. Ahbab, S. Ahmed, T. Ahmed, and M. H. Kabir, “Improved speech emotion recognition in bengali language using deep learning,” in *2023 26th International Conference on Computer and Information Technology (ICCIT)*, IEEE, 2023, pp. 1–6.
- [46] M. M. Billah, M. L. Sarker, and M. Akhand, “Kbes: A dataset for realistic bangla speech emotion recognition with intensity level,” *Data in Brief*, vol. 51, p. 109741, 2023.
- [47] R. Das and T. D. Singh, “Multimodal sentiment analysis: A survey of methods, trends, and challenges,” *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–38, 2023.
- [48] P. Deb, “Multitask audio analysis for emotion, gender, and speaker recognition in bangla speech comparing features and models,” in *2023 International Conference on Advanced Computing Technologies and Applications (ICACTA)*, IEEE, 2023, pp. 1–6.
- [49] K. T. Elahi, T. B. Rahman, S. Shahriar, S. Sarker, S. K. S. Joy, and F. M. Shah, “Explainable multimodal sentiment analysis on bengali memes,” in *2023 26th International Conference on Computer and Information Technology (ICCIT)*, IEEE, 2023, pp. 1–6.
- [50] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, “Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions,” *Information Fusion*, vol. 91, pp. 424–444, 2023.
- [51] A. Ghorbanali and M. K. Sohrabi, “A comprehensive survey on deep learning-based approaches for multimodal sentiment analysis,” *Artificial Intelligence Review*, vol. 56, no. Suppl 1, pp. 1479–1512, 2023.
- [52] C. Huang, J. Zhang, X. Wu, Y. Wang, M. Li, and X. Huang, “Tefna: Text-centered fusion network with crossmodal attention for multimodal sentiment analysis,” *Knowledge-Based Systems*, vol. 269, p. 110502, 2023.

- [53] M. M. H. Jibon, D. M. Alam, and M. S. Rahman, “Banglabeats: A comprehensive dataset of bengali songs for music genre classification tasks,” in *2023 26th International Conference on Computer and Information Technology (ICCIT)*, IEEE, 2023, pp. 1–6.
- [54] M. N. R. Khan, K. I. J. Tuli, M. S. Salsabil, S. R. Reza, F. M. Shah, and S. K. S. Joy, “Bengali music genre classification using wavenet,” in *2023 26th International Conference on Computer and Information Technology (ICCIT)*, IEEE, 2023, pp. 1–6.
- [55] 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 (Big-Comp)*, IEEE, 2023, pp. 219–226.
- [56] L. Sun, Z. Lian, B. Liu, and J. Tao, “Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis,” *IEEE Transactions on Affective Computing*, 2023.
- [57] Z. S. Taheri, A. C. Roy, and A. Kabir, “Bemofusionnet: A deep learning approach for multimodal emotion classification in bangla social media posts,” in *2023 26th International Conference on Computer and Information Technology (ICCIT)*, IEEE, 2023, pp. 1–6.
- [58] K. K. Veni, S. I. A. Lathif, K. S. Kumar, T. Subha, S. A. Shifani, and D. Nesakumar, “A novel emotion recognition model based on speech processing,” in *2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, IEEE, 2023, pp. 1–7.
- [59] F. Wang, S. Tian, L. Yu, *et al.*, “Tedt: Transformer-based encoding–decoding translation network for multimodal sentiment analysis,” *Cognitive Computation*, vol. 15, no. 1, pp. 289–303, 2023.
- [60] C. Zhu, M. Chen, S. Zhang, *et al.*, “Skeafn: Sentiment knowledge enhanced attention fusion network for multimodal sentiment analysis,” *Information Fusion*, vol. 100, p. 101 958, 2023.
- [61] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, and X. Kong, “Multimodal sentiment analysis based on fusion methods: A survey,” *Information Fusion*, vol. 95, pp. 306–325, 2023.
- [62] Z. Liu, B. Zhou, D. Chu, Y. Sun, and L. Meng, “Modality translation-based multimodal sentiment analysis under uncertain missing modalities,” *Information Fusion*, vol. 101, p. 101 973, 2024.
- [63] Q. Lu, X. Sun, Z. Gao, Y. Long, J. Feng, and H. Zhang, “Coordinated-joint translation fusion framework with sentiment-interactive graph convolutional networks for multimodal sentiment analysis,” *Information Processing & Management*, vol. 61, no. 1, p. 103 538, 2024.
- [64] Y. Zeng, W. Yan, S. Mai, and H. Hu, “Disentanglement translation network for multimodal sentiment analysis,” *Information Fusion*, vol. 102, p. 102 031, 2024.
- [65] “*the world factbook*”. *www.cia.gov. central intelligence agency. archived from the original on 26 january 2021. retrieved 21 february 2018*. <https://www.cia.gov/the-world-factbook/about/archives/>.

- [66] *Bengali at ethnologue (27th ed., 2024)*.
- [67] <https://medium.com/@jdkiptoon/feature-selection-in-machine-learning-20417d052b80>, <https://medium.com/@jdkiptoon/feature-selection-in-machine-learning-20417d052b80>.
- [68] <https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270>, <https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270>.
- [69] <https://www.shedloadofcode.com/blog/eight-ways-to-perform-feature-selection-with-scikit-learn>, <https://www.shedloadofcode.com/blog/eight-ways-to-perform-feature-selection-with-scikit-learn>.
- [70] *Multimodal ai use cases - google cloud*, <https://cloud.google.com/use-cases/multimodal-ai>.
- [71] *What is multimodal fusion*, <https://www.educative.io/answers/what-is-multimodal-fusion>.

# Appendix I

## Grid Search Results

### AdaBoost

From table 5.1 we can see the details of grid search results for the AdaBoost model. The search runs for 30 combinations including five different learning rates (0.001,0.01, 0.1, 0.5,1) and six different estimators (50,100,200,300,400,500). From the table, it is evident that the best mean test score of 0.623 is achieved by the combination where the learning rate is 0.5 and the estimator is 50. We have used this combination for our model.

Table 5.1: Grid Search Results for AdaBoost

mean fittime	std fit time	mean score time	std score time	param lear ning <i>rate</i>	param n estim ators	split0 test score	split1 test score	split2 test score	split3 test score	split4 test score	mean test score	std test score	rank test score
0.53516001	0.0041827	0.0131628	0.0011408	0.001	50	0.512974	0.51	0.516	0.5	0.514	0.510594	0.00564	24
1.05214395	0.009544	0.020780	0.000422	0.001	100	0.512974	0.51	0.516	0.5	0.5	0.507794	0.006641	28
2.29170341	0.273979	0.047380	0.011874	0.001	200	0.512974	0.51	0.522	0.5	0.492	0.507394	0.01042	30
3.8317119	0.584764	0.074546	0.021137	0.001	300	0.51497	0.51	0.522	0.5	0.492	0.507794	0.010662	29
4.42836098	0.383018	0.088900	0.024848	0.001	400	0.51497	0.516	0.522	0.5	0.492	0.508994	0.011168	25
5.67721805	0.401606	0.121032	0.028391	0.001	500	0.51497	0.52	0.524	0.492	0.49	0.508194	0.014342	26
0.66771059	0.091247	0.016322	0.004043	0.01	50	0.51497	0.52	0.524	0.492	0.49	0.508194	0.014342	26
1.04929866	0.007449	0.021033	0.000733	0.01	100	0.55489	0.52	0.536	0.516	0.52	0.529378	0.014489	23
2.28791580	0.277807	0.044305	0.008314	0.01	200	0.58483	0.566	0.552	0.518	0.544	0.552966	0.022312	22
3.3527344	0.235053	0.065731	0.010847	0.01	300	0.60678	0.574	0.562	0.522	0.562	0.565357	0.027179	21
4.56764025	0.444953	0.076984	0.003493	0.01	400	0.602794	0.596	0.584	0.526	0.568	0.575358	0.027364	20
5.65770111	0.413872	0.108338	0.019929	0.01	500	0.60479	0.626	0.596	0.54	0.576	0.588558	0.029108	18
0.52750420	0.005096	0.011927	0.000551	0.1	50	0.59481	0.622	0.596	0.538	0.586	0.587362	0.027453	19
1.24418673	0.163462	0.027964	0.006184	0.1	100	0.610778	0.63	0.628	0.552	0.608	0.605755	0.028293	12
2.32934799	0.246332	0.044642	0.012607	0.1	200	0.618762	0.632	0.652	0.576	0.604	0.616552	0.025699	8
3.35945196	0.212828	0.063916	0.012852	0.1	300	0.616766	0.628	0.66	0.572	0.618	0.618953	0.028203	5
4.57816753	0.463870	0.077525	0.003729	0.1	400	0.608782	0.636	0.654	0.57	0.622	0.618156	0.028372	6
5.59118523	0.367960	0.119633	0.028550	0.1	500	0.610778	0.626	0.66	0.578	0.624	0.619755	0.02646	3
0.52195706	0.005520	0.011725	0.000430	<b>0.5</b>	<b>50</b>	0.606786	0.616	0.648	0.594	0.652	<b>0.623357</b>	0.022883	<b>1</b>
1.14568953	0.155110	0.026764	0.006015	0.5	100	0.61477	0.61	0.64	0.57	0.63	0.612954	0.023996	11
2.18654136	0.187899	0.039258	0.002168	0.5	200	0.608782	0.616	0.66	0.582	0.622	0.617756	0.025158	7
3.53096604	0.471876	0.058241	0.001378	0.5	300	0.616766	0.626	0.662	0.58	0.622	0.621353	0.026097	2
4.36852221	0.391281	0.076216	0.00211	0.5	400	0.620758	0.624	0.654	0.582	0.616	0.619351	0.022937	4



Table 5.1: Grid Search Results for AdaBoost

mean fittime	std fit time	mean score time	std score time	param lear ning <i>rate</i>	param n estim ators	split0 test score	split1 test score	split2 test score	split3 test score	split4 test score	mean test score	std test score	rank test score
5.65231189	0.408061	0.108845	0.030540	0.5	500	0.61477	0.61	0.642	0.582	0.624	0.614554	0.019612	9
0.59535288	0.086701	0.014697	0.003730	1	50	0.60479	0.586	0.608	0.568	0.62	0.597358	0.018287	16
1.05365014	0.014138	0.020707	0.000199	1	100	0.628743	0.578	0.628	0.55	0.632	0.603348	0.033353	13
2.28624539	0.240922	0.045973	0.013510	1	200	0.602794	0.612	0.652	0.564	0.638	0.613758	0.030494	10
3.3857394	0.274768	0.065569	0.014620	1	300	0.602794	0.598	0.632	0.552	0.608	0.598558	0.026052	15
4.57170224	0.458404	0.074174	0.001583	1	400	0.598802	0.588	0.618	0.552	0.594	0.590160	0.021565	17
5.64107708	0.362571	0.107954	0.029257	1	500	0.602794	0.592	0.634	0.57	0.612	0.602158	0.021206	14