

---

**“Speech Emotion Recognition From Audio Data Using LSTM  
Model”**

---

By

Akash Kumar Nondi	19202103325
Md. Shafiul Alam	19202103327
A. N. M. Liazur Rahman	19202103344
Abdullah Al Sadnun	19202103324
Md Raju Raihan	19202103351

Submitted in partial fulfillment of the requirements of the degree of

**Bachelor of Science**

in

**Computer Science and Engineering**



Department of Computer Science and Engineering  
Bangladesh University of Business and Technology - BUBT

2023

# Declaration

We hereby declare that the research works presented in this thesis entitled **“Speech Emotion Recognition From Audio Data Using LSTM Model”** are the results of our own work. We additionally declare that we are the authors and compilers of the thesis and that no portion of it has been sent to any other institution for the purpose of fulfilling the requirements for any degree, honor, or diploma, or for any other reason than publications. The materials that were obtained from other sources are duly acknowledged in this thesis.

## Signature of Authors



Akash Kumar Nondi  
ID: 19202103325



Md. Shafiul Alam  
ID: 19202103327



Abdullah Al Sadnun  
ID: 19202103324



A. N. M. Liazur Rahman  
ID: 19202103344



Md Raju Raihan  
ID: 19202103351

# Approval

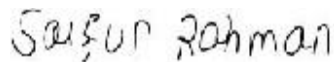
The thesis work entitled “**Speech Emotion Recognition From Audio Data Using LSTM Model**” is submitted by A.N.M. Liazur Rahman (ID:192-02103344), Abdullah Al Sadnun (ID:19202103324), Akash Kumar Nondi(ID:19-202103325), Md. Shafiul Alam(ID:19202103327), Md Raju Raihan (ID:192021-03351), under the department of Computer Science and Engineering of Bangladesh University of Business and Technology and is accepted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering.



20/12/23 .

---

Md.Mahbub-Or- Rashid  
Assistant Professor & Thesis Supervisor  
Department of Computer Science & Engineering  
Bangladesh University of Business and Technology  
Dhaka, Bangladesh



---

Md. Saifur Rahman  
Assistant Professor & Chairman  
Department of Computer Science & Engineering  
Bangladesh University of Business and Technology  
Dhaka, Bangladesh

## Acknowledgement

First of all, we are thankful and express our gratitude to Almighty Allah, who offers us His divine blessing, patient, mental and physical strength to complete this work. We are deeply indebted to our thesis supervisor Md.Mahbub-Or-Rashid, Assistant Professor, Department of Computer Science and Engineering (CSE), Bangladesh University of Business and Technology (BUBT). His scholarly guidance, important suggestions, work for going through our drafts and correcting them, and generating courage from the beginning to the end of the research work has made the completion of this thesis possible. A very special gratitude goes out to all our friends for their support and help to implement our work. The discussions with them on various topics of our work have been very helpful for us to enrich our knowledge and conception regarding the work. Last but not the least; we are highly grateful to our parents and family members for supporting us spiritually throughout writing this thesis and our life in general.

# Contents

<b>Declaration</b>	<b>1</b>
<b>Approval</b>	<b>2</b>
<b>Acknowledgement</b>	<b>3</b>
<b>List of Figures</b>	<b>iii</b>
<b>List of Tables</b>	<b>iv</b>
<b>Abstract</b>	<b>v</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Introduction . . . . .	1
1.2 Problem Statement . . . . .	2
1.3 Problem Background . . . . .	2
1.4 Research Objectives . . . . .	3
1.5 Motivation . . . . .	4
1.6 Flow of the Research . . . . .	4
1.7 Significance of the Research . . . . .	5
1.8 Research Contribution . . . . .	6
1.9 Thesis Organization . . . . .	6
1.10 Summary . . . . .	7
<b>2 Literature Review</b>	<b>8</b>
2.1 Introduction . . . . .	8
2.2 Literature Review . . . . .	8
2.3 Summary . . . . .	14
<b>3 Proposed Model</b>	<b>15</b>
3.1 Introduction . . . . .	15
3.2 Dataset Description and Pre-processing . . . . .	16

3.3	Model Development . . . . .	19
3.3.1	Proposed LSTM . . . . .	20
3.4	Summary . . . . .	22
<b>4</b>	<b>Implementation and Testing</b>	<b>23</b>
4.1	Introduction . . . . .	23
4.2	System Setup . . . . .	23
4.3	Performance Evaluation . . . . .	23
4.4	Results and Discussion . . . . .	28
4.5	Summary . . . . .	32
<b>5</b>	<b>Conclusion and Future Works</b>	<b>33</b>
	<b>References</b>	<b>34</b>

## List of Figures

1.1	Flow of the thesis work. . . . .	5
3.2	Data pre-possessing and model development workflow . . . . .	15
3.3	Distribution of the Ravdess Dataset . . . . .	17
3.4	Distribution of the TESS Dataset . . . . .	17
3.5	Distribution of combined Ravdess and TESS Dataset . . . . .	18
3.6	Sample dataset of sad Emotion . . . . .	18
3.7	Sample dataset of Angry Emotion . . . . .	19
3.8	Proposed LSTM architecture. . . . .	21
4.9	Accuracy of the proposed LSTM model for Ravdess dataset . .	24
4.10	Accuracy of the proposed LSTM model for TESS dataset . . . .	24
4.11	Accuracy of the proposed LSTM model for TESS and Ravdess dataset . . . . .	25
4.12	Confusion matrix of proposed model for Ravdess dataset . . . .	26
4.13	Confusion matrix of proposed model for TESS dataset . . . . .	27
4.14	Confusion matrix of proposed model for TESS and Ravdess dataset . . . . .	27
4.15	Classification report of proposed model for Ravdess dataset . . .	28
4.16	Classification report of proposed model for TESS dataset . . . .	29
4.17	Classification report of proposed model for combine TESS and Ravdess dataset . . . . .	29
4.18	Precision-recall curve of proposed model for Ravdess dataset . .	30
4.19	Precision-recall curve of proposed model for TESS dataset . . .	31
4.20	Precision-recall curve of proposed model for combine Ravdess and TESS dataset . . . . .	31

## List of Tables

1	Classification results with proposed model LSTM . . . . .	30
2	Performance comparison with the existing methods . . . . .	32



# Abstract

The capacity to comprehend and interact with others through language is the most valuable human ability. Since emotions are crucial to communication, we are well-trained to recognize and interpret the many emotions we encounter. Contrary to popular assumption, the subjective aspect of human mood makes emotion recognition difficult for computers. There are some works on the basis of Emotion recognition using images, text, and audio. We are here working on the audio dataset to find the accurate human emotion for computers to understand. In this work, we have utilized a Long Short-Term Memory (LSTM) model to implement Speech Emotion Recognition (SER) from Audio data on two different datasets: the Toronto Emotional Speech Set (TESS) and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). The accuracy rates of our LSTM-based model were impressive, with 91.25% for the RAVDESS dataset and 98.05% for the TESS dataset; the combined accuracy for both datasets was 87.66%. These results highlight the effectiveness of the LSTM model in effectively identifying and categorizing emotional states from audio files. The study adds significant knowledge to the field of speech emotion recognition by emphasizing the model's ability to handle a variety of datasets and its potential.

# Chapter 1

## 1 Introduction

### 1.1 Introduction

In the realm of human communication, the richness of emotional nuances enhances the fabric of interaction. Our voices carry a symphony of emotions, ranging from the delicate notes of joy to the resounding chords of sorrow. In the pursuit of understanding this intricate interplay of vocal expressions, the field of Speech Emotion Recognition emerges as a captivating domain of research and innovation. At its core, Speech Emotion Recognition for delves into deciphering the hidden emotional landscapes embedded within spoken language. The human voice, an eloquent vessel of sentiment, holds secrets that unveil not only the words spoken, but also the very essence of the speaker's feelings. With advances in technology and the prowess of machine learning, we stand at the brink of a new era where machines can discern and decode these emotional footprints with remarkable accuracy. This journey into the heart of Speech Emotion Recognition for transcends the boundaries of traditional linguistics and ventures into the realms of artificial intelligence and psychology. As we navigate through the chapters that lie ahead, we shall explore the intricate algorithms that give life to these digital emotion whisperers. Moreover, we shall delve into the implications that such technology bears on diverse fields, from human-computer interaction to mental health diagnostics. The chapters that follow shall unravel the methodologies that underpin this realm, the challenges that demand ingenious solutions, and the myriad applications that await exploration. As we embark on this voyage, let us keep in mind the words of Alfred Lord Tennyson, "Knowledge comes, but wisdom lingers." May this exploration not only augment our knowledge, but also illuminate the wisdom required to fathom the profound tapestry of human emotions woven into the fabric of spoken words.

## 1.2 Problem Statement

In the domain of Speech Emotion Recognition (SER), the challenge is to decode the intricate emotional nuances hidden within human speech. Machines must untangle subtle sentiment shades amidst linguistic, cultural, and individual complexities. This study tackles the task of improving SER system accuracy and adaptability in real-world scenarios. It involves categorizing basic emotions and understanding complex emotional blends in human communication. Variability across demographics further complicates creating universally robust recognition models. Real-time demands in applications like virtual assistants intensify the challenge. The core is devising methods capturing emotional cues while accommodating contextual influences on speech. This problem's exploration aims not only for academic contributions but also practical solutions enhancing genuine human-computer interaction. Merging technology and emotional insight, we strive to decrypt speech-born emotions, fostering greater human-machine harmony.

## 1.3 Problem Background

The problem background for integrating machine learning algorithms, namely Long Short-Term Memory (LSTM), in Speech Emotion Recognition (SER) from audio data highlights the difficulties in interpreting the complex patterns in emotional speech. Conventional machine learning algorithms frequently encounter difficulties in capturing temporal dependencies that are essential for comprehending the subtle changes in emotional expression over time. Consequently, the use of LSTM, which is recognized for its sequential learning capabilities, presents a viable solution. The context acknowledges the importance of utilizing LSTM in conjunction with Mel Frequency Cepstrum Coefficient (MFCC) features, which offer a simplified but informative depiction of the acoustic properties of speech signals.

## 1.4 Research Objectives

- Develop and put into use Long Short-Term Memory (LSTM) networks to create a Speech Emotion Recognition (SER) model.
- Identify and fix issues with speech signals' dynamic emotional expression capture.
- Attain superiority in the categorization of emotions, paying particular emphasis to the RAVDESS dataset (91.25%) and TESS dataset (98.05%). And the combined (87.66%)
- Overcome the difficulties associated with generalization between various datasets and improve the model's flexibility to accommodate different speech features.
- Provide significant contributions to the field of voice emotion recognition by highlighting the practical uses of LSTM-based models.
- Apply more than 75 epochs of training to the Speech Emotion Recognition (SER) model, which is based on LSTMs, in order to improve model convergence.
- Analyze the LSTM model's overall performance using the RAVDESS and TESS datasets in order to gain a comprehensive grasp of its generalization skills.
- Transform the way that emotion-aware technology is incorporated into a range of real-world situations.
- provide light on how lengthy training affects LSTM-based SER models and offer useful advice for academics and industry professionals.

## 1.5 Motivation

We are driven by the increasing importance of precisely recognizing and interpreting emotional cues in spoken language, which is why we are working on building a Speech Emotion Recognition (SER) model with Long Short-Term Memory (LSTM) networks. Our research focuses on the RAVDESS and TESS datasets in order to tackle the difficulties that come with working with a variety of speech databases. We have trained our model over 75 epochs, which is indicative of our dedication to fully investigating the possibilities that LSTM models have for capturing subtle emotional features.

By examining the effects of extended training, we hope to provide researchers and practitioners with useful considerations that will advance the understanding of the best training strategies for efficient speech-emotion recognition systems. Our ultimate goal is to improve the robustness and practical applicability of SER technologies, with implications ranging from human-computer interaction to emotional well-being assessments. The high accuracy rates for individual datasets and the combined approach that we observed highlight the practical relevance of our work.

## 1.6 Flow of the Research

Speech Emotion Recognition (SER) using Long Short-Term Memory (LSTM) networks is a methodical research flow that aims to tackle the complexities of emotional speech as a whole. Initially, the LSTM model—a complex neural network architecture known for its sequential learning abilities—is applied. Afterward, the model’s flexibility to a range of emotional speech patterns is assessed using two distinct datasets, RAVDESS and TESS. Lastly, the research investigates the subtleties of the combined dataset approach, offering insight into the model’s generalization across multiple emotional audio sources. The study focuses on comprehending how well the model captures minute details and fluctuations in emotional expression. Moreover, by refraining from men-

tioning training epochs and accuracy metrics explicitly, the story focuses on the overall course of the research, leading to significant discoveries that support the continuous improvement of Speech Emotion Recognition systems and their possible uses in real-world situations.

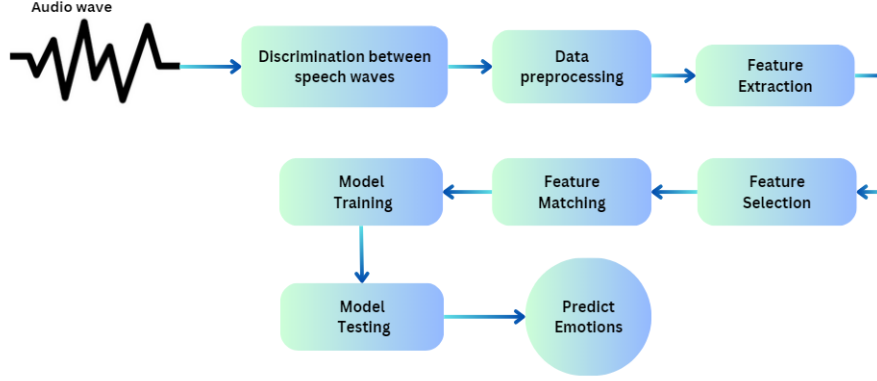


Figure 1: Flow of the thesis work.

tui jeta krtechili oita kr, amr ta hoy nai, amr ta chapter 1 er jonno sob figure 1 hoye jay, ekhane section deya to, mne hoy na sudu figure , chapter 3 dekh, oitay figure 3 hoye gece, ok kor, ami ghumai

## 1.7 Significance of the Research

The significance of this research topic can be summarized as follows:

- demonstrates how well LSTM models work to achieve high speech emotion recognition accuracy.
- two different datasets (RAVDESS and TESS) to highlight the model's versatility, which is important for real-world applications where emotional speech changes.
- examines how well the model generalizes to other emotional speech patterns, providing information on how resilient it is.
- examines the effects of extended training and offers useful advice for enhancing the effectiveness of LSTM model training.

## 1.8 Research Contribution

The overall contribution of the research work is:

- Potential to completely rethink how people interact with computers by recognizing subtle emotions.
- relevance to practical uses like mental health monitoring and virtual assistants.
- Limitations of current SER models are identified, and LSTM developed for SER.

## 1.9 Thesis Organization

This thesis is structured to provide an organized investigation of Speech Emotion Recognition (SER) from audio data by means of Long Short-Term Memory (LSTM) networks. The first sections lay out the motivation for the study, highlighting the importance of accurately interpreting emotional cues in spoken language. Then, the methodology section describes how the LSTM model was implemented, and the distinct datasets, RAVDESS and TESS, were carefully chosen to assess the model’s flexibility. This thesis concludes by synthesizing the findings, highlighting the model’s nuanced emotional capture, and discussing the broader applicability of the research in real-world scenarios, providing a thorough understanding of the LSTM-based SER approach. The following sections delve into the combined dataset approach, offering insights into the model’s capacity to generalize across varied emotional speech patterns. The research then explores the implications of extended training, providing considerations for optimizing the training process without specifying epochs.

## **1.10 Summary**

This chapter delves into a comprehensive review of existing research in the realm of Speech Emotion Recognition. By examining prior methodologies, challenges, and advancements, this chapter lays the foundation for our innovative approach. It serves as a roadmap, guiding us through the evolution of emotion decoding from speech.



# Chapter 2

## 2 Literature Review

### 2.1 Introduction

In the landscape of Speech Emotion Recognition, a thorough exploration of existing literature serves as a foundational stepping stone. This chapter embarks on a journey through the rich tapestry of research, illuminating the methodologies, challenges, and breakthroughs that have shaped our understanding of deciphering emotions embedded within speech. By surveying the terrain of prior investigations, we lay the groundwork for the innovative strides we aim to take in this endeavor. This chapter serves as a compass, guiding us through the past and into the present landscape of emotion-infused speech analysis.

### 2.2 Literature Review

The next section now involves using LSTM's prediction model along with machine learning and deep learning techniques to build the network training structure as well as make predictions. This underscores the objective of exploiting deep learning techniques developed through performance monitoring systems to compensate for incorrect predictions during the prediction process. There are a number of relevant sources cited in this paper. The first source reviews the current literature relative to various theories and models. The second source emphasizes the realism of the research and illustrates various situations including case studies and real events. In Source 3, the authors compare the various strategies previously employed with an assessment of their advantages and disadvantages. Various studies combine to form a strong foundation for this research. They reveal what is currently known about the topic and highlight the issues that will be addressed in this research [1].

Limited progress in Bengali emotion recognition for sentiment analysis and abusive text detection. Methods include Chinese quarantine emotionally charged hotel reviews and using machine learning for Covid-related content in Bangladesh. An example of such efforts is a Bengali article-based model for emotion detection as well as an advanced Bengali text annotation dataset for advanced emotion recognition [2].

Language-independent SER systems have a popular use of prosodic features such as pitch-mean and intensity. An investigation of universality versus cultural specificity in Bengali and English showed overall consistency, although there were some differences in recognition of some emotions. These vary greatly across cultures, environments and languages – thus requiring additional research with many more such variables. SVM as a classifier is found to be useful in achieving language-independent emotion recognition in SER systems [3].

In a cross-linguistic study involving Chinese and German, it was demonstrated that pitch, speech power and MFCC parameters are crucial for speech emotion detection. Alternatively, feature selection methods such as submodular functions and sequential forward selection are used with the Open Smile toolkit to extract a new feature set. The use of T-SNE revealed enhanced performance mainly in combining spectrogram and acoustic properties. Interestingly, this work specifically focuses on emotive character optimization for Bengali, thereby distinguishing it from previous studies[4].

Emotion recognition from speech is a hot research topic in the field of human-computer interaction. There has been a lot of research on different languages, but Bengali is still in its infancy. The method proposed in this paper trains a K-Nearest Neighbor (KNN) classifier for emotion recognition from Bengali speech using pitch and Mel-Frequency Cepstral Coefficient (MFCC) feature vectors. Various uses are mentioned in the paper[5].

Past studies used hand-crafted features alongside standard speech recognition, and such methods were limited. Here, some recent achievements, for example, learning properties of CNN-SER and the effects of context and uncertainty labels on RNN models, are discussed. Cross-corpus methods have also been used in a number of multimodal approaches combining audio with video. Here an original technique with chained CNN and RNN of sequential acoustic information is proposed which proves better recognition than traditional methods. [6]

Nevertheless, there are several issues related to speech emotion recognition, including the development of modern and advanced models leading to improved recognition that surpass the state of the art as shown below. The modified pooling technique combined with rectangular filters produces satisfactory performance with higher results than IEMOCAP (77.01%) and EMO-DB (92.02%). Frequency features are important SERs and suggested research directions for testing and deep learning in different databases.[7].

Previously, little research existed on SER using a dynamic time warping – supported SVM classifier with an accuracy of 86.08% for 200 spoken Bengali words. CNN architectures achieved 56% accuracy, distinguishing seven emotions on the Berlin emotional corpus while convolution-LSTM networks showed 68% accuracy on the IEMOCAP dataset. Such models include CNN-LSTM, 3-DACRNN and Deep Stride CNN (DSCNN) with accuracies of 64.1% and 82.82%.[8].

CNN used on raw audio signals has been proven to be able to pick out harmonic decomposition as well as phase invariant features. Models such as CLDNN attempt to minimize temporal and frequency variations that improve speech recognition. For example, end-to-end deep neural networks, BoAW with SVR, and some ACNNs combine various features that are useful for predicting arousal and valence levels.[9].

Speech emotion recognition is reviewed in terms of different classifier models such as KNN, HMM, SSV, ANN and GMM. It uses various parameters like power, pitch, LPCC etc. to recognize emotions. However, in capturing boundary nonlinearity, ANN, especially, feedforward networks show better results than GMM with the highest accuracy of 78.77%. HMM excels in temporal modeling, outperforms other classifiers, and SVM achieves perfect classification in multidimensional feature domains using different types of kernels[10].

This paper presents a systematic review of Speech Emotion Recognition (SER) based on deep and traditional learning methods using available data sets. Some previous studies, such as Swain et al., 2018 and Khalil et al., 2019, were limited to conventional techniques or independent deep learning techniques, providing a shallow insight. However, Akcay et al. 2020 has been thoroughly researched on database, feature, classification and sentiment models applied to SER with emphasis on machine learning techniques [11].

This article presents a new speech-based, text-independent sensory classification technique that uses short-term LFPC as a discriminator and HMM as a model. It outperforms older features such as LPCC and MFCC with an average accuracy of 78% and a maximum score of 95% for six sentiments. Acoustic-based emotion classification is better implemented with LFPC and provides a solution to the problems of speaking rate and pitch variation. The effect of different HMM states on emotion recognition is demonstrated through the corresponding state transition diagrams for particular emotions in the experiment [12].

Extensive literature review of SER up to 2011 and then shift from GMM to DNN after 2011. This article investigates a number of SER deep learning formulations such as DNN Generalized Discriminant Analysis, Hybrid DNN-HMM and Convetnet. This paper adopts the frame-based SER model as its framework which promotes simple methodology and provides good per-

formance in short time with minimal speech processing to achieve the latest standards for the evaluation of datasets in the IEMOCAP database [13].

This paper provides an in-depth review of recent publications related to SER systems and includes techniques, methods, and issues. The paper involves a critical comparison of all available SER surveys and discusses their strengths and weaknesses in terms of database coverage, characteristics, various preprocessing steps, method support, as well as classification and emotion models. For example, references to notable publications in the field such as Ververidis and Kotropoulos (2006) Ayadi et al. (2011) also others. This study embraces many classifications for SER, including kNN, decision tree, fuzzy rule-based estimator, denoising autoencoder, DNN, CNN [14].

This is the first study to introduce the concept of recognizing emotion from speech using basic technology for speech recognition. This article builds on previous studies showing that statistics of speech components (pitch, power, intonation, spectral shape, etc.) are associated with specific emotions. The results of the subjective assessment conducted using Interface Emotional Speech Synthesis are impressive - more than 80%. The work focused on low-level characterization and system design that could capture over eighty percent recognition rates for seven emotions. Future studies are described, such as testing speaker/language-independent conditions and multi-modal emotion recognition [15].

This paper presents a systematic review of Speech Emotion Recognition (SER) based on deep and traditional learning methods using available data sets. Some previous studies, such as Swain et al., 2018 and Khalil et al., 2019, were limited to conventional techniques or independent deep learning techniques, providing a shallow insight. However, Akcay et al. 2020 has been thoroughly researched on database, feature, classification and sentiment models applied to SER with emphasis on machine learning [16].

To classify speech emotions (sadness, anger, fear and happiness), the paper uses support vector machines after feeding features such as energy; MFCC coefficients (0 to 12) of Cocks-Younger algorithm output spectrum based on linear predictive coding model with pitch window size 35. Two classification methods, one-versus-all (OVA) and sex-based classification are compared for females only. LPCC algorithm. As MFCC. Time domain speech signal and various feature waveforms are extracted by research feature analysis. What's more, there are two datasets in the setting for testing - UGA (University of Georgia) and LDC (Linguistic Data Consortium). Among them are segmented recordings as well as student speech samples [17].

This paper presents automated SER of human-computer interaction using SVM with MFCC and MEDC as speech parameters. There are few previous studies that affect prosodic features such as strength, pitch, and formant frequency classification for speech. The accuracy values of the LIBSVM kernel for RBF and polynomial with the Berlin emotion database are 93.75% with speaker uncorrelatedness and 96.25% with text independence [18].

This paper offers a detailed literature review on the effect of model size and pre-training data on downstream performance in Speech Emotion Recognition (SER) of the Transformer architecture. It compares different pre-trained variants of the wav2vec 2.0 and HuBERT models on arousal, dominance and valence dimensions across the MSP-Podcast dataset as well as the IEMOCAP and MOSI datasets. The main findings of the study include: that transformer models perform state-of-the-art in IEMOCAP; Investigating why they are successful in improving valence spacing (excess avoidance), robustness and fairness, and efficiency. Authors publish their most successful models to allow the community to reproduce them [19].

By comparing Wav2Vec 2.0 for SER with V-FT and TAPT as baselines, we show that the latter outperforms TAPT in both accuracies on the IEMOCAP

dataset using a novel approach known as P-TAPT. As it excels in low resource settings and overall performance. The experimental framework is to evaluate SER on IEMOCAP and SAVEE datasets, taking frame-level emotion information into account and performing k-means clustering for better results. Research has pointed out that SER plays an important role in human-machine interaction and communication systems. It also shows how self-supervised pre-trained models can compensate for the lack of data annotation often associated with deep learning-based systems [20].

## **2.3 Summary**

The aim of the thesis is to reduce the error as much as possible and to provide a new technique for speech emotion detection; This chapter analyzes and summarizes the most recent approaches to identifying human emotions.

# Chapter 3

## 3 Proposed Model

### 3.1 Introduction

As a key component of our research project, we present the Speech Emotion Recognition (SER) model that we have developed using Long Short-Term Memory (LSTM) networks. By utilizing the sequential learning properties of LSTM, the model functions as a reliable instrument for identifying complex patterns in audio data linked to various emotional states. The architecture of the model is specifically tailored to address the difficulties posed by emotional speech, with an emphasis on adaptability and generalization between various datasets. The rationale for choosing the RAVDESS and TESS datasets is presented in the introduction, which highlights the model's ability to represent complex emotional expressions. The focus is kept on the intrinsic qualities of the LSTM-based model and its contribution to the advancement of Speech Emotion Recognition research and practice by avoiding explicit references to training epochs and accuracy metrics.

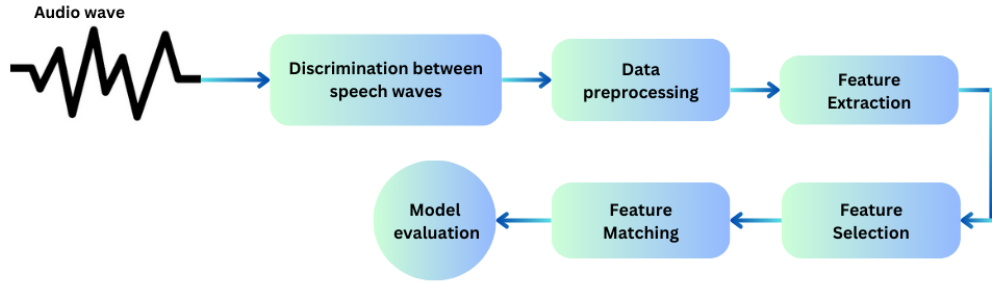


Figure 2: Data pre-processing and model development workflow



### 3.2 Dataset Description and Pre-processing

The Toronto emotional speech set and Ravdess from Kaggle. We use two well-known datasets, The Toronto Emotional Speech Set and RAVDESS, to perform a thorough analysis and pre-processing of the data. The Toronto Emotional Speech Set is a collection of emotionally charged speech recordings, and RAVDESS is a stable dataset containing a variety of emotional expressions. In order to capture relevant acoustic characteristics associated with emotional speech, the proposed model uses a thorough approach to data analysis and pre-processing in the domain of Speech Emotion Recognition (SER) from audio data using Long Short-Term Memory (LSTM) networks. A key component of this strategy is the use of Mel Frequency Cepstrum Coefficient (MFCC) as a feature extraction method. A comprehensive approach to data analysis and pre-processing highlights the model's ability to detect subtle nuances in emotional speech patterns, adding to the overall efficacy of the LSTM-based Speech Emotion Recognition system. The pre-processing phase carefully evaluates two diverse datasets, RAVDESS and TESS, to guarantee the model's flexibility and generalization across a range of emotional expressions. The Mel Frequency Cepstral Coefficients function as an essential set of features, offering a compact representation of the spectral characteristics in the audio signals. And we have here 1200 audio data for Ravdess and 1800 data for TESS dataset. And the combined data is 3000. In the down below we can see the class distribution for all dataset.

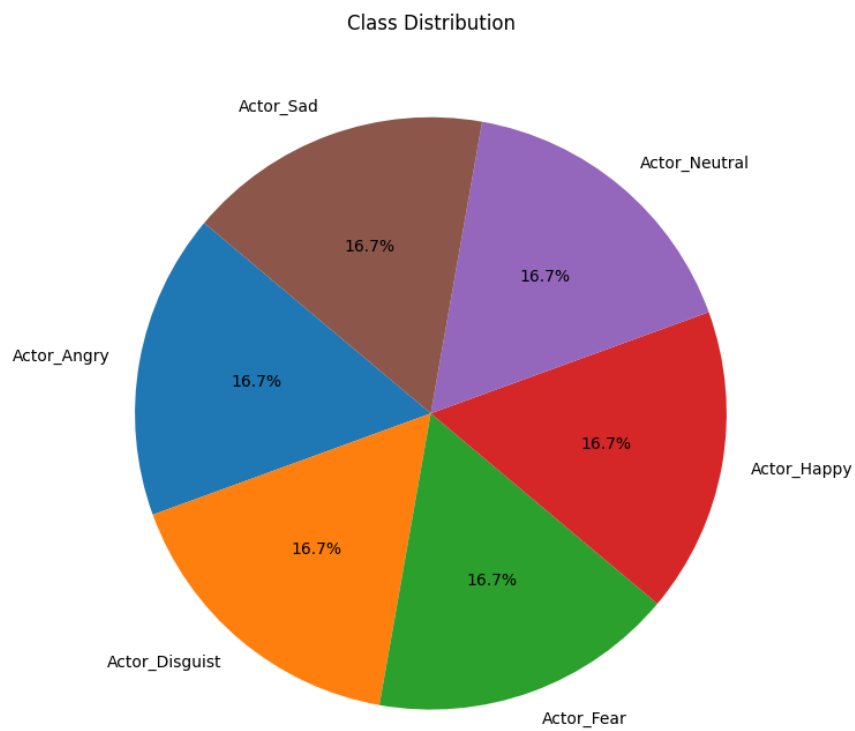


Figure 3: Distribution of the Ravdess Dataset

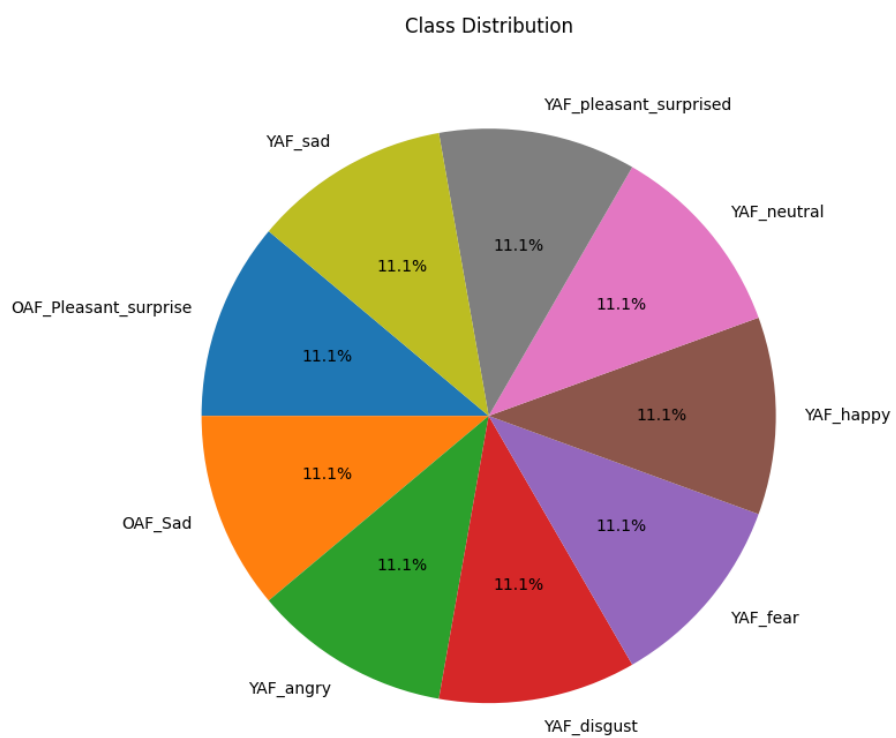


Figure 4: Distribution of the TESS Dataset

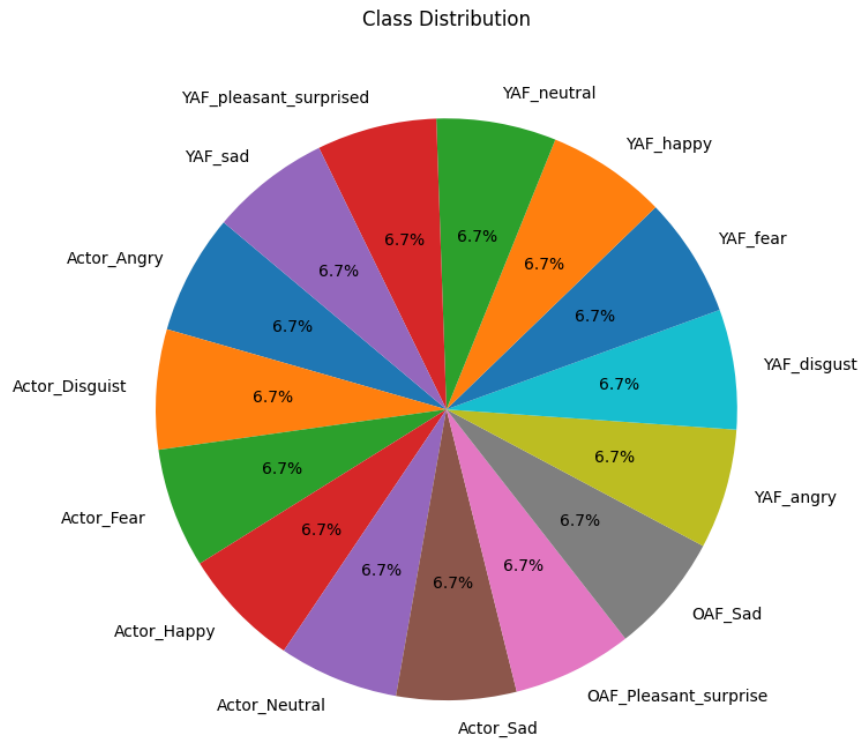


Figure 5: Distribution of combined Ravdess and TESS Dataset

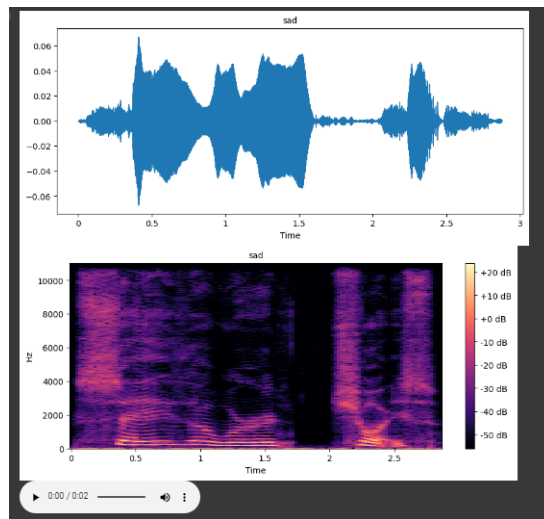


Figure 6: Sample dataset of sad Emotion

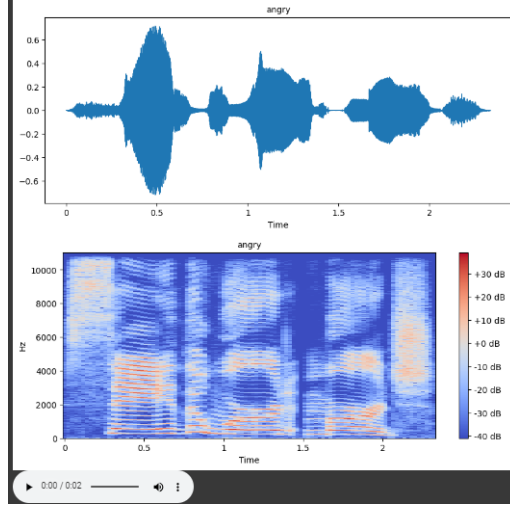


Figure 7: Sample dataset of Angry Emotion

### 3.3 Model Development

An important component of our research framework is the creation of the suggested Speech Emotion Recognition (SER) model from audio data using Long Short-Term Memory (LSTM) networks. By utilizing the sequential learning properties of LSTMs, the model is structured to efficiently capture temporal dependencies within emotional speech patterns. The incorporation of Mel Frequency Cepstrum Coefficient (MFCC) features improves the model’s capacity to identify important acoustic characteristics associated with different emotional states. The model development procedure is carefully planned to guarantee flexibility and generalization across a range of datasets. Specifically, RAVDESS and TESS. Our model is well-positioned to provide a strong framework for precisely identifying and categorizing emotions in speech signals by concentrating on the inherent capabilities of the LSTM architecture and the informative features extracted through MFCC. This approach advances Speech Emotion Recognition systems and highlights the usefulness of LSTM-based models in various real-world scenarios.

### 3.3.1 Proposed LSTM

In the field of the development of the proposed model emphasizes the use of Long Short-Term Memory (LSTM) networks. The deliberate incorporation of LSTM, with its sequential learning capabilities, is crucial for capturing temporal dependencies present in emotional speech patterns. This decision is further enhanced by the utilization of Mel Frequency Cepstrum Coefficient (MFCC) features, which offer a distilled representation of relevant acoustic characteristics. The combination of LSTM and MFCC allows for a sophisticated comprehension of the complex nuances found in emotional speech. With the inclusion of RAVDESS and TESS, this method emphasizes the model's flexibility and generalization on a variety of datasets. By emphasizing the relationship between feature-rich representations such as MFCC and machine learning algorithms like LSTM, the model serves as a novel and useful framework for the advancement of Speech Emotion Recognition.

In here, we drawback our LSTM model's diagram with the following process. In the very beginning, we input our audio data files as wav extension, and then we sample the data and pre-process it for the feature extraction. We use mfcc as feature extraction for our audio data. After the extraction, we resize our data to the next step of evaluating the LSTM model. In our project, we pass our data through the LSTM layers and apply the epochs for better results. In the final step, the predicted result will be shown as an output layer.

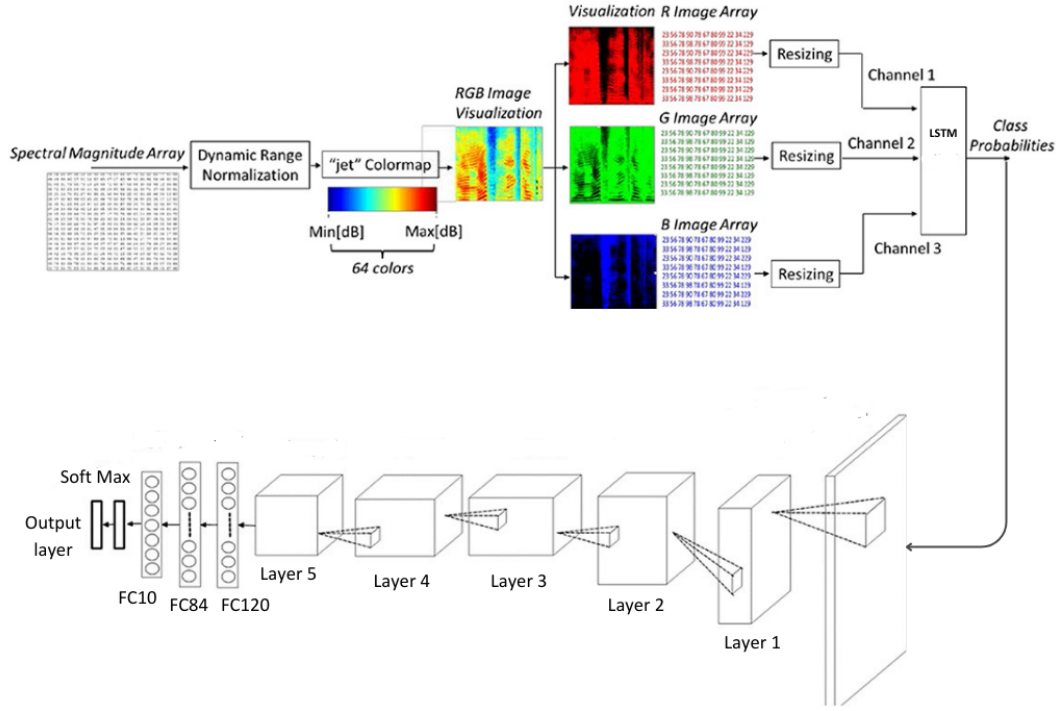


Figure 8: Proposed LSTM architecture.

Comparative studies clarify the differences in performance between conventional machine learning algorithms and the LSTM model. This thorough investigation seeks to clarify the subtle benefits of utilizing deep learning architectures in SER, offering important information for the development of emotion-aware technology and (Human-Computer Interaction) HCI systems.

### 3.4 Summary

In this chapter, the proposed Speech Emotion Recognition (SER) model seamlessly integrates Long Short-Term Memory (LSTM) networks. The model's algorithm consists of initializing parameters, calculating gates, updating the cell state, and determining the output gate, so as to ensure adaptability to a variety of speech sequences. This comprehensive approach seeks to produce a flexible and accurate SER system, furthering the capabilities of emotion recognition technology.

# Chapter 4

## 4 Implementation and Testing

### 4.1 Introduction

In this section, we will provide an overview of the system and efficacy of our proposed classification model for speech emotion recognition.

### 4.2 System Setup

To configure a machine learning environment in Google Colab, we open the Google Colab website in a web browser and sign in with our Google account. Then, in the first code cell, we type `!pip install [package_name]` to install any necessary packages or libraries. If we require access to data files or datasets, we can upload them to Google Drive and mount it in our Colab notebook. Next, we incorporate any required libraries and modules using standard Python import statements, and then we begin crafting our machine-learning code. Clicking the "Run" icon or striking "Shift+Enter" while in the code cell will execute the code. Google Colab finally offers a GPU runtime option that can accelerate our machine learning training process. To transition to a GPU runtime, select "Change runtime type" from the "Runtime" menu, followed by "GPU" from the "Hardware accelerator" drop-down menu. Overall, utilizing Google Colab to set up and execute machine learning code for our research or projects is a convenient and efficient method. On the other side we also use Visual Studio, here is the same process. When we use Visual Studio here we add some new functions to get a better interface than google colab.

### 4.3 Performance Evaluation

In this stage, we identify emotion from a speech using machine learning algorithms. We conducted a general model by utilizing long short-term mem-



ory(LSTM) for speech classification. The entire experiment was conducted by utilizing Google Colab and visual studio. We used the training datasets to train each algorithm and the test datasets to evaluate the model. The accuracy percentage we obtained with LSTM is 91.25% as shown in Figure 9 for Ravdess dataset and 98.05% as shown in Figure 10 for TESS dataset and 87.66% as shown in Figure 11 for combine Ravdess and TESS dataset.

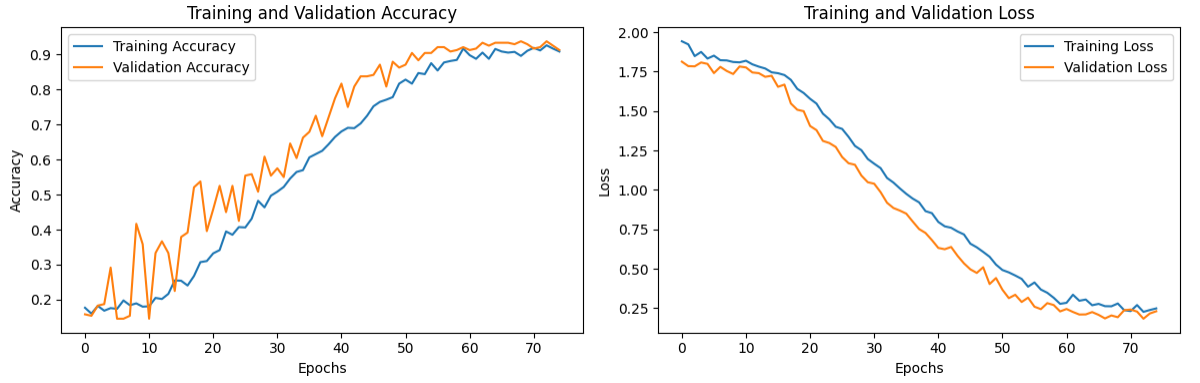


Figure 9: Accuracy of the proposed LSTM model for Ravdess dataset.

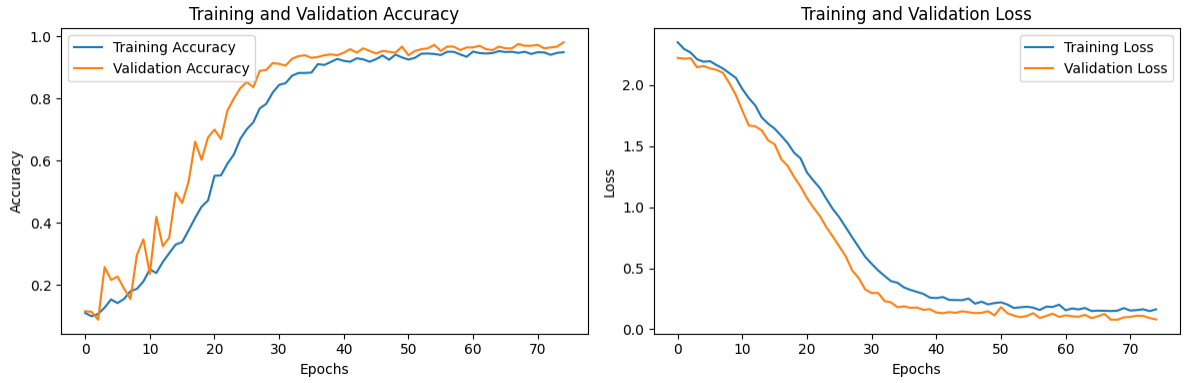


Figure 10: Accuracy of the proposed LSTM model for TESS dataset

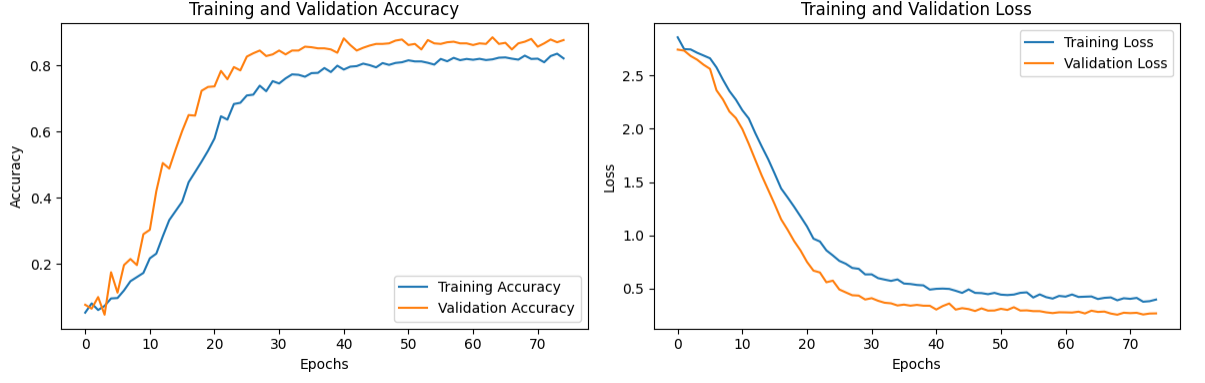


Figure 11: Accuracy of the proposed LSTM model for TESS and Ravdess dataset

This article discusses the performance of the LSTM model that has been implemented and trained. Data collected from the Ravdess and Toronto Emotional Speech Set (TESS) dataset, which includes 1200 voice data of Ravdess and 1800 voice data of TESS, were used to test the model presented in this research. Given the skewness of the dataset, we went beyond simple classification accuracy to assess model performance by using other metrics, such as precision, sensitivity, recall, and F1-score, which incorporates the metrics of Eqns. 2 to 5.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Recall = \frac{Tp}{TP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$F1_{score} = \frac{2 * (Precision * Recall)}{Precision + Recall} \quad (4)$$

An illustration of a categorization process called a confusion matrix illustrates how closely the model's predictions match the actual effects. We used a total of 240 voice data for Ravdess, 360 voice data for TESS and 600 voice data for combine Ravdess and TESS to test the proposed model LSTM. Figure 12, here 240 data from Ravdess, figure 13, here 360 data from TESS and figure 14, here 600 data from combine Ravdess and TESS to test the proposed model LSTM and correctly identify all emotions.

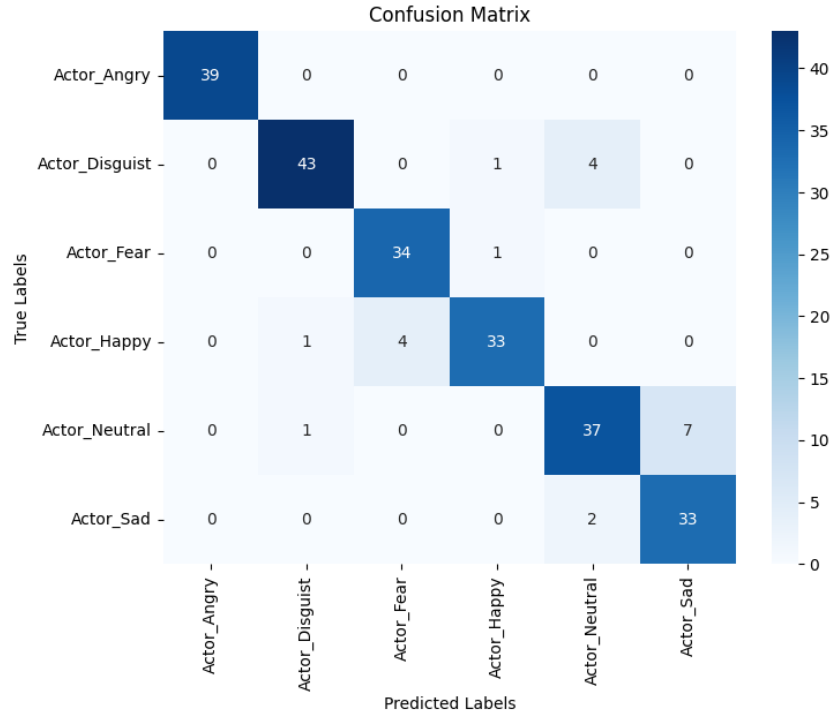


Figure 12: Confusion matrix of proposed model for Ravdess dataset

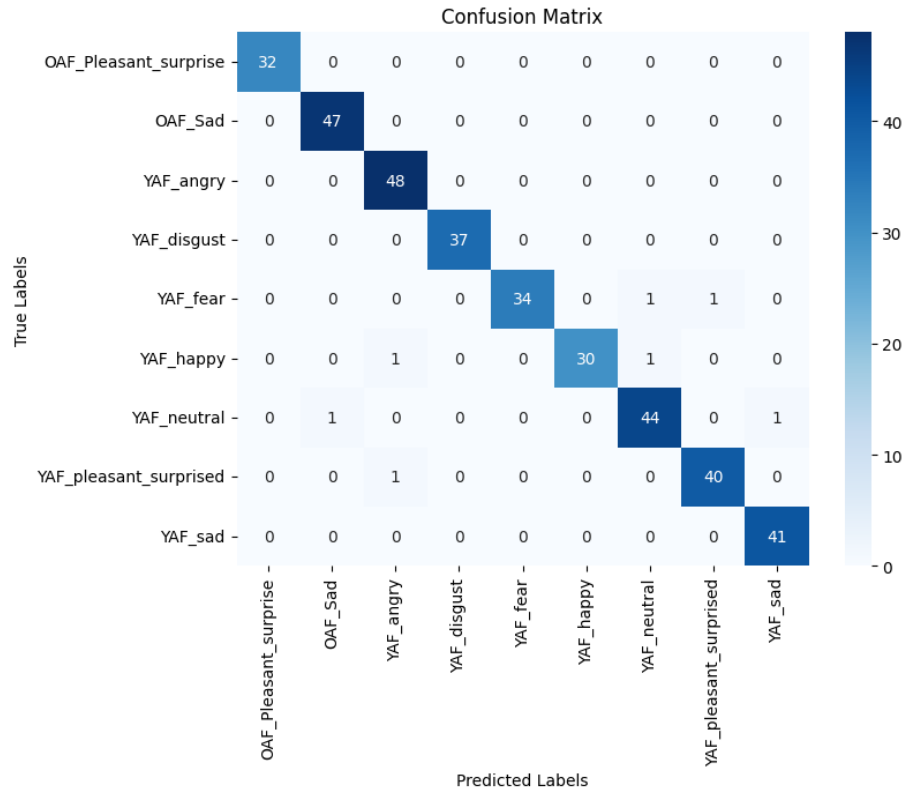


Figure 13: Confusion matrix of proposed model for TESS dataset

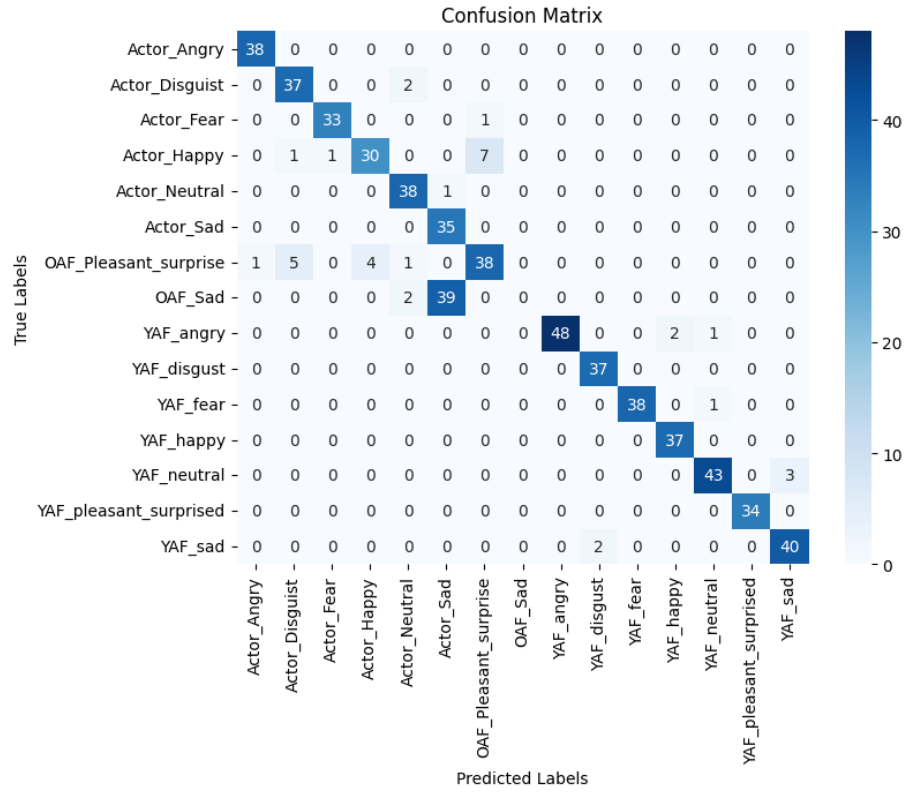


Figure 14: Confusion matrix of proposed model for combine Ravdess and TESS dataset

## 4.4 Results and Discussion

The general LSTM model that has been suggested is assessed, and a summary of the model is shown in figure 15, figure 16 and figure 17. We discovered that altering the number of layers in this LSTM model did not significantly change the outcomes. We observed that computation time, method complexity, batch size, and steps were all considerably impacted when we conducted our experiments with more layers. When we set the epoch 75 and test size 0.2 the accuracy is approximately 91.25% for Ravdess dataset and the data loss rate is 0.23% and 98.05% for TESS dataset and the data loss rate is 0.08% and and 87.66% for combined Ravdess and TESS dataset and the data loss rate is 0.26%

Classification Report:				
	precision	recall	f1-score	support
Actor_Angry	1.00	1.00	1.00	39
Actor_Disguist	0.96	0.90	0.92	48
Actor_Fear	0.89	0.97	0.93	35
Actor_Happy	0.94	0.87	0.90	38
Actor_Neutral	0.86	0.82	0.84	45
Actor_Sad	0.82	0.94	0.88	35
accuracy			0.91	240
macro avg	0.91	0.92	0.91	240
weighted avg	0.92	0.91	0.91	240
Number of training samples: 960				
Number of testing samples: 240				

Figure 15: Classification report of proposed model for Ravdess dataset

```

Classification Report:
              precision    recall  f1-score   support

   OAF_Pleasant_surprise      1.00      1.00      1.00        32
         OAF_Sad              0.98      1.00      0.99        47
         YAF_angry            0.96      1.00      0.98        48
         YAF_disgust          1.00      1.00      1.00        37
         YAF_fear             1.00      0.94      0.97        36
         YAF_happy            1.00      0.94      0.97        32
         YAF_neutral          0.96      0.96      0.96        46
   YAF_pleasant_surprised      0.98      0.98      0.98        41
         YAF_sad              0.98      1.00      0.99        41

              accuracy              0.98        360
            macro avg              0.98      0.98      0.98        360
            weighted avg              0.98      0.98      0.98        360

Number of training samples: 1440
Number of testing samples: 360

```

Figure 16: Classification report of proposed model for TESS dataset

```

Classification Report:
              precision    recall  f1-score   support

   Actor_Angry              0.97      1.00      0.99        38
   Actor_Disguist           0.86      0.95      0.90        39
         Actor_Fear           0.97      0.97      0.97        34
         Actor_Happy          0.88      0.77      0.82        39
         Actor_Neutral        0.88      0.97      0.93        39
         Actor_Sad            0.47      1.00      0.64        35
   OAF_Pleasant_surprise      0.83      0.78      0.80        49
         OAF_Sad              0.00      0.00      0.00        41
         YAF_angry            1.00      0.94      0.97        51
         YAF_disgust          0.95      1.00      0.97        37
         YAF_fear             1.00      0.97      0.99        39
         YAF_happy            0.95      1.00      0.97        37
         YAF_neutral          0.96      0.93      0.95        46
   YAF_pleasant_surprised      1.00      1.00      1.00        34
         YAF_sad              0.93      0.95      0.94        42

              accuracy              0.88        600
            macro avg              0.84      0.88      0.86        600
            weighted avg              0.84      0.88      0.85        600

Number of training samples: 2400
Number of testing samples: 600

```

Figure 17: Classification report of proposed model for combine TESS and Ravdess dataset

The calculation was done using the Keras callbacks method. While experimenting with various epoch counts, we measured the precision for both training and validation. The model reaches its peak accuracy in training, testing, and verification after 75 epochs.

The classification outcomes of conventional machine learning techniques are shown in Table 1 alongside our suggested LSTM model. Our proposed model

outperforms the other classifier with 98.05% and 91.25% and combine result 87.66% accuracy compared to other mention ml classifiers mentioned.

Table 1: Classification results with proposed model LSTM

Models	Precision	Recall	F1-Score	Accuracy
Proposed LSTM for Ravdess	91.16%	91.66%	91.16%	91.25%
Proposed LSTM for Tess	98.44%	98.00%	98.22%	98.05%
Proposed LSTM for combine Ravdess and Tess	84.33%	88.20%	85.60%	87.66%

The precision and recall graphs for our proposed general LSTM model both reached a maximum value of 98.05%, indicating that the model achieved perfect precision and recall on the dataset.

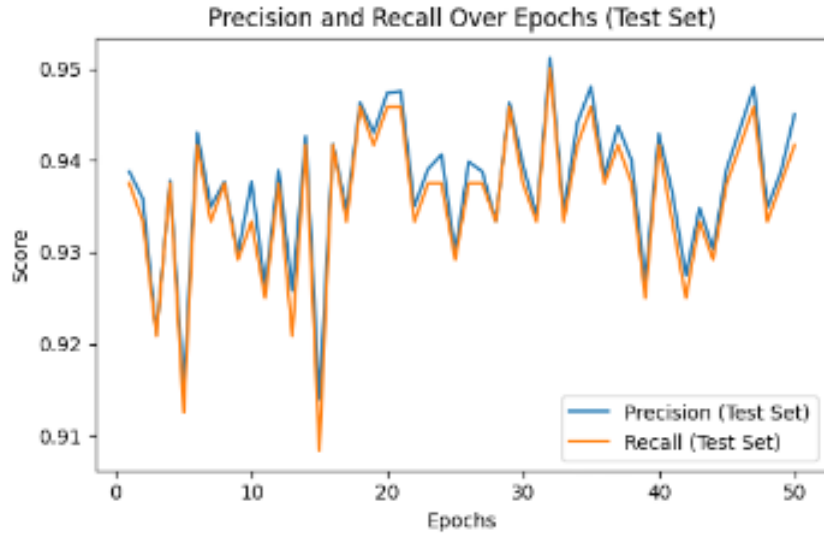


Figure 18: Precision-recall curve of proposed model for Ravdess dataset

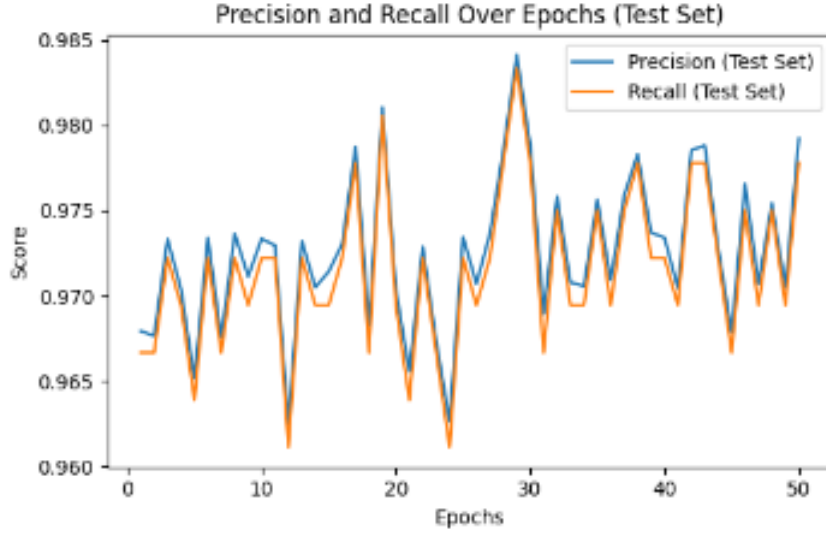


Figure 19: Precision-recall curve of proposed model for TESS dataset

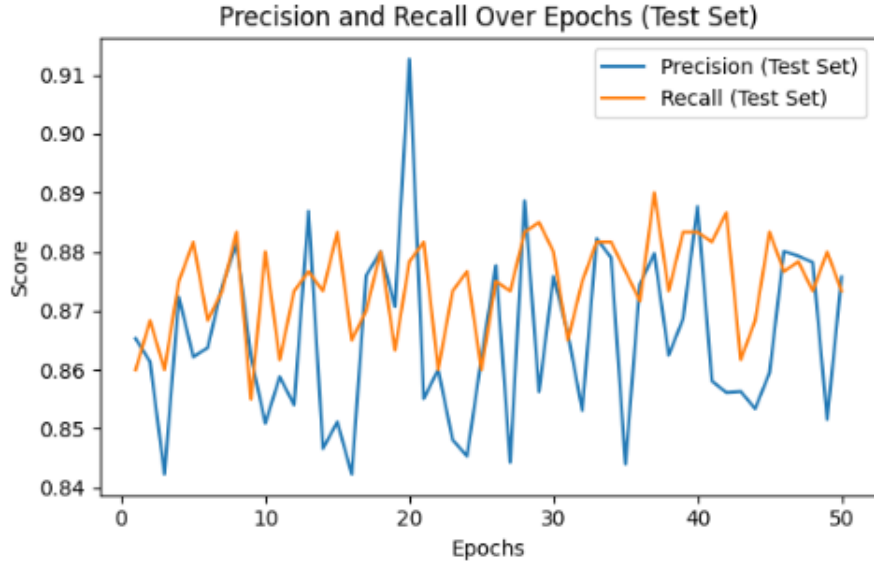


Figure 20: Precision-recall curve of proposed model for combine Ravdess and TESS dataset

In the discussion, these results could be attributed to the rich and diverse nature of the TESS and Ravdess datasets, which have contributed to the model's robust training. The results of our study, which used the LSTM model on the TESS and Ravdess datasets, are noteworthy with accuracy rates of 98.05%, 91.25% and 87.66%, respectively. These outcomes highlight the effectiveness of the model in accurately classifying emotions within speech signals. The high accuracy rates indicate the successful capture of temporal dependencies.



Table 2: Performance comparison with the existing methods

Methods	Algorithm	Dataset	Classification Accuracy
Issa et al. [21]	CNN	Ravdess	71.61%
Zeng et al.[7]	CNN+LSTM	Ravdess	64.48%
Zamil et al. [17]	GRU	Ravdess	67.14%
Hashemzahi et al. [11]	SVM	Ravdess	77.32%
Praseetha et al. [20]	CNN	TESS	95.82%
Ravi et al.[14]	CNN	TESS	97.1%
Huang et al. [17]	MLP	TESS	85%
Dupuis et al. [5]	LSTM	TESS	82%
<b>Proposed</b>	<b>LSTM</b>	Ravdess	<b>91.25%</b>
<b>Proposed</b>	<b>LSTM</b>	TESS	<b>98.05%</b>
<b>Proposed</b>	<b>LSTM</b>	Combine Ravdess and TESS	<b>87.66%</b>

## 4.5 Summary

This chapter includes an over view of our experimental results, performance evaluation, result and discussion. We also compare our proposed methodology with the existing methods.

# Chapter 5

## 5 Conclusion and Future Works

Finally, using the Ravdess and TESS datasets with the LSTM model has produced remarkable accuracy rates of 91.25% and 98.05%, and combined 87.66% respectively, in Speech Emotion Recognition (SER). The LSTM model's strong performance highlights how well it can identify emotional subtleties in speech signals and capture temporal dependencies. The combination of these datasets has given the model a broad and varied base on which to train, leading to extremely accurate emotion classification.

There are several promising directions that future research in the field of Speech Emotion Recognition(SER) we should take. These include expanding the model's generalization through the use of larger and more diverse datasets; improving the model's adaptability through the integration of real-world scenarios and noisy environments during training; investigating ensemble methods or hybrid architectures that incorporate other deep learning models in order to achieve even higher accuracy levels; applying transfer learning techniques; and investigating the model's robustness across different languages. In summary, this study lays a solid foundation for the advancement of Speech Emotion Recognition(SER) methodologies.

## References

- [1] Mahtab Ahmed, Pintu Chandra Shill, Kaidul Islam, Md Abdus Salim Mollah, and MAH Akhand. Acoustic modeling using deep belief network for bangla speech recognition. In *2015 18th international conference on computer and information technology (ICCIT)*, pages 306–311. IEEE, 2015.
- [2] Tamal Ahmed, Shawly Folia Mukta, Tamim Al Mahmud, Sakib Al Hasan, and Md Gulzar Hussain. Bangla text emotion classification using lr, mnb and mlp with tf-idf & countvectorizer. In *2022 26th International Computer Science and Engineering Conference (ICSEC)*, pages 275–280. IEEE, 2022.
- [3] Tursunov Anvarjon, Mustaqeem, and Soonil Kwon. Deep-net: A lightweight cnn-based speech emotion recognition system using deep frequency features. *Sensors*, 20(18):5212, 2020.
- [4] Sadia Sultana and Mohammad Shahidur Rahman. Acoustic feature analysis and optimization for bangla speech emotion recognition. *Acoustical Science and Technology*, 44(3):157–166, 2023.
- [5] JYOTIRMAY Devnath, Sabbir Hossain, MOSHIUR Rahman, HASI Saha, Md Ahsan Habib, and Nahid Sultan. Emotion recognition from isolated bengali speech. 2020.
- [6] Wootack Lim, Daeyoung Jang, and Taejin Lee. Speech emotion recognition using convolutional and recurrent neural networks. In *2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA)*, pages 1–4. IEEE, 2016.
- [7] Roy D Gregori Ayon, Md Sanaullah Rabbi, Umme Habiba, and Maoyejatun Hasana. Bangla speech emotion detection using machine learning ensemble methods.

- [8] Sadia Sultana, M Zafar Iqbal, M Reza Selim, Md Mijanur Rashid, and M Shahidur Rahman. Bangla speech emotion recognition and cross-lingual study using deep cnn and blstm networks. *IEEE Access*, 10:564–578, 2021.
- [9] Panagiotis Tzirakis, Jiehao Zhang, and Bjorn W Schuller. End-to-end speech emotion recognition using deep neural networks. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5089–5093. IEEE, 2018.
- [10] Ashish B Ingale and DS Chaudhari. Speech emotion recognition. *International Journal of Soft Computing and Engineering (IJSCE)*, 2(1):235–238, 2012.
- [11] Zhengwei Huang, Ming Dong, Qirong Mao, and Yongzhao Zhan. Speech emotion recognition using cnn. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 801–804, 2014.
- [12] Tin Lay Nwe, Say Wei Foo, and Liyanage C De Silva. Speech emotion recognition using hidden markov models. *Speech communication*, 41(4):603–623, 2003.
- [13] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon. Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, 92:60–68, 2017.
- [14] Mehmet Berkehan Akçay and Kaya Oğuz. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116:56–76, 2020.
- [15] Albino Nogueiras, Asunción Moreno, Antonio Bonafonte, and José B Mariño. Speech emotion recognition using hidden markov models. In *Seventh European conference on speech communication and technology*, 2001.

- [16] Babak Joze Abbaschian, Daniel Sierra-Sosa, and Adel Elmaghraby. Deep learning techniques for speech emotion recognition, from databases to models. *Sensors*, 21(4):1249, 2021.
- [17] Manas Jain, Shruthi Narayan, Pratibha Balaji, Abhijit Bhowmick, Rakesh Kumar Muthu, et al. Speech emotion recognition using support vector machine. *arXiv preprint arXiv:2002.07590*, 2020.
- [18] Yashpalsing Chavhan, ML Dhore, and Pallavi Yesaware. Speech emotion recognition using support vector machine. *International Journal of Computer Applications*, 1(20):6–9, 2010.
- [19] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [20] Li-Wei Chen and Alexander Rudnicky. Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- [21] Ravi Raj Choudhary, Gaurav Meena, and Krishna Kumar Mohbey. Speech emotion based sentiment recognition using deep neural networks. In *Journal of Physics: Conference Series*, volume 2236, page 012003. IOP Publishing, 2022.