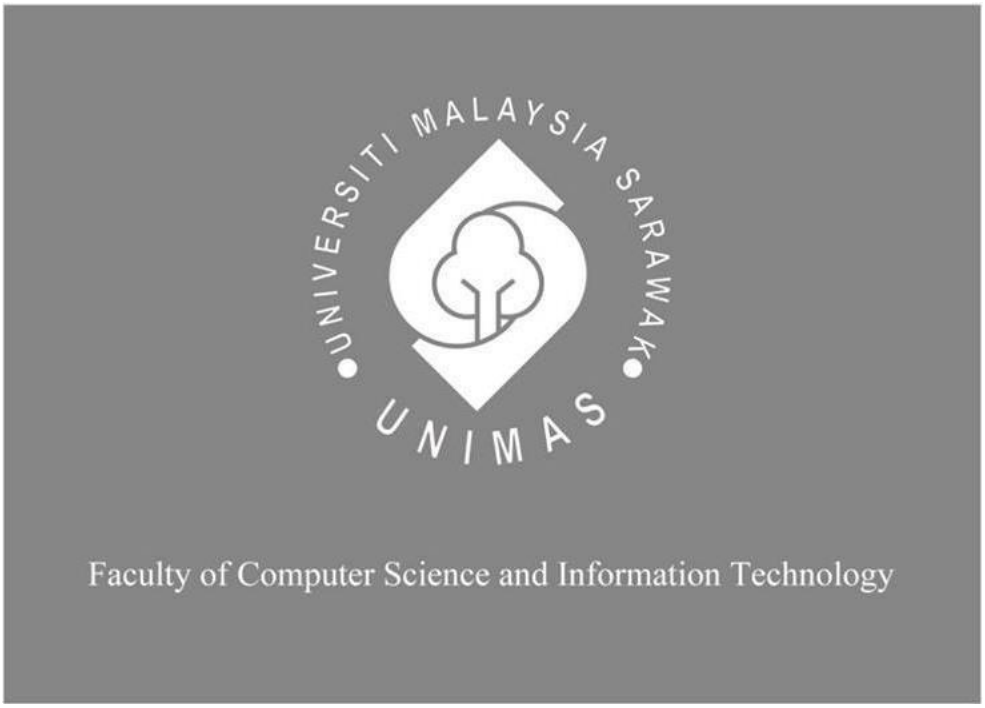


Emotion Recognition from Adult Speech Using Machine Learning: A Study Based on the CREMA-D Dataset

Eric Anak Naweam

Bachelor Of Computer Science and Technology (Hons)

Information System

2024

Emotion Recognition from Adult Speech Using Machine Learning: A Study Based on the CREMA-D Dataset

ERIC ANAK NAWEAM

This project is submitted in partial fulfilment of the requirements for the degree of Bachelor of Computer Science with Honours (Information System)

Faculty of Computer Science and Information Technology
UNIVERSITI MALAYSIA SARAWAK

2024

UNIVERSITI MALAYSIA SARAWAK

THESIS STATUS ENDORSEMENT FORM

TITLE **Emotion Recognition from Adult Speech Using Machine Learning: A Based on the CREMA-D Dataset**

ACADEMIC SESSION: **2024/25**

ERIC ANAK NAWEAM (79325)

(CAPITAL LETTERS)

hereby agree that this Thesis* shall be kept at the Centre for Academic Information Services, Universiti Malaysia Sarawak, subject to the following terms and conditions:

1. The Thesis is solely owned by Universiti Malaysia Sarawak
2. The Centre for Academic Information Services is given full rights to produce copies for educational purposes only
3. The Centre for Academic Information Services is given full rights to do digitization in order to develop local content database
4. The Centre for Academic Information Services is given full rights to produce copies of this Thesis as part of its exchange item program between Higher Learning Institutions [or for the purpose of interlibrary loan between HLI]
5. ** Please tick (√)

☐	CONFIDENTIAL	(Contains classified information bounded by the OFFICIAL SECRETS ACT 1972)
☐	RESTRICTED	(Contains restricted information as dictated by the body or organization where the research was conducted)
☑	UNRESTRICTED	

(AUTHOR'S SIGNATURE)

Validated by

(SUPERVISOR'S SIGNATURE)

Permanent Address

Kampung Tringgus Rabak Rotan, 94000 Bau, Sarawak

Date: 15/7/2025

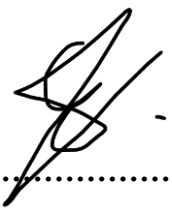
Date: 21 July 2025

Note * Thesis refers to PhD, Master, and Bachelor Degree

** For Confidential or Restricted materials, please attach relevant documents from relevant organizations / authorities

DECLARATION

I, Eric Anak Naweam, 79325, hereby declare that the project entitled “Emotion Recognition from Adult Speech Using Machine Learning: A Study Based on the CREMA-D Dataset” is my original work and no part of this work has been written by others on my behalf. I have not plagiarized works from others or other sources in completing the project, except where references or acknowledgments are provided. I have diligently adhered to the academic guidelines and ethical standards set by the Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak (UNIMAS), in completing this project.



18/06/2025

…….. …….

ERIC ANAK NAWEAM **DATE**

Acknowledgement

First and foremost, I am deeply grateful to the Lord for granting me the strength, wisdom, and perseverance to complete this final year project. Without His guidance and blessings, this accomplishment would not have been possible.

I would also like to express my heartfelt gratitude to my supervisor, Ts. Dr Sarah Flora Anak Samson Juan, for her invaluable guidance, encouragement and constructive feedback throughout this project. Her expertise and support have been instrumental in shaping the direction and success of this research.

I also wish to extend my thanks to the Faculty of Computer Science and Information Technology (FCSIT), Universiti Malaysia Sarawak (UNIMAS), for providing the resources and platform to pursue this study. The knowledge and skills I gained during my academic journey have been essential in completing this final year project.

Special thanks to my family and friends for their unwavering support, understanding and encouragement during challenging times. Their belief in me has been my greatest motivation in completing this project. Thank you once again to everyone who played a role in this journey.

Table of Content

List of Figures

List of Tables

List of Equations

Abstract

Speech Emotion Recognition (SER) is the ability of a machine to understand and interpret human emotions. This ability has emerged as an important component in enhancing human-computer interaction (HCI), enabling systems to understand and respond to human emotions effectively. This research addresses several challenges in SER, including variability in emotional expression due to factors such as background noise, accents, and linguistic diversity, the difficulty in selecting suitable deep learning architectures and the existing datasets used for training and testing emotion recognition models often do not reflect the variety of emotional expressions found in real-world speech. This study focuses on analysing deep learning techniques commonly used for speech emotion recognition, including CNNs, LSTMs, Transformers, and hybrid models. A hybrid model integrating CNN and LSTM architectures is proposed, leveraging the CREMA-D dataset for training and testing. Feature extraction techniques such as Mel spectrograms are employed, and the model's performance is evaluated using metrics including accuracy, precision, recall, and F1-score. Python, TensorFlow, and Keras are used to implement the models, with development conducted on platforms like Visual Studio Code. Through comprehensive evaluation and analysis, the proposed 2D CNN-LSTM model achieved an accuracy of 59.1%, surpassing the baseline 2D CNN and demonstrating enhanced recognition of emotional states. However, despite this improvement, the overall accuracy remains relatively low compared to previous SER studies. Key challenges include difficulty in detecting subtle emotions caused by dataset imbalance. Future work suggests the further study on integrating transformer-based architectures, advanced data augmentation techniques and cross-validation methods to improve model generalization and performance.

Keywords: Speech Emotion Recognition, Deep Learning, Convolutional Neural Networks, Long Short-Term Memory (LSTM) Networks, CREMA-D, Feature Extraction, Cross Validation.

Abstrak

Pengecaman Emosi Melalui Pertuturan atau lebih dikenali sebagai *Speech Emotion Recognition (SER)* merujuk kepada kemampuan sebuah system atau mesin untuk memahami dan menginterpretasikan emosi manusia. Kemampuan ini telah muncul sebagai suatu komponen penting dalam meningkatkan keupayaan interaksi antara manusia dan komputer atau lebih dikenali sebagai *Human-Computer Interaction (HCI)*, yang memungkinkan sistem untuk memahami dan merespons kepada emosi manusia secara efektif. Kajian ini membincangkan beberapa cabaran yang dihadapi dalam Pengecaman Emosi Melalui Pertuturan, termasuk kepelbagaian cara dalam mengekspresikan emosi termasuk faktor-faktor seperti suara latar belakang, aksen, dan kepelbagaian bahasa. Selain itu, kesusahan dalam memilih teknik pembelajaran mendalam atau *Deep Learning (DL)* yang sesuai dan set data yang sedia ada yang digunakan untuk melatih dan pengujian model Pengecaman Emosi Melalui Pertuturan ini sering kali tidak mencerminkan kepelbagaian ekspresi emosi yang ditemukan dalam pertuturan dunia sebenar. Kajian ini mengfokuskan kepada menganalisis teknik pembelajaran mendalam yang biasa digunakan untuk Pengecaman Emosi Melalui Pertuturan, termasuk rangkaian saraf perlingkaran atau *Convolutional Neural Networks (CNNs),* Rangkaian Memori Jangka Pendek Panjang atau *Long Short-Term Memory Networks (LSTMs),* Transformer, dan model gabungan. Sebuah model gabungan yang mengintegrasikan teknik *CNN*s dan *LSTM*s diusulkan, dengan memanfaatkan dataset *CREMA-D* untuk melatih dan pengujian. Teknik pengekstrakan ciri suara seperti *Mel Spectrogram* turut digunakan, dan keupayaan model untuk mengecam emosi dinilai menggunakan metrik seperti ketepatan *(accuracy)*, ketepatan *(precision), recall,* dan nilai F1. Python, TensorFlow, dan Keras digunakan untuk mengimplementasikan model ini, dengan pembinaan model ini yang dilakukan pada platform seperti Visual Studio Code. Melalui penilaian dan analisis menyeluruh, model 2D CNN-LSTM yang diusulkan mencapai ketepatan

sebanyak 59.1%, melebihi model asas 2D CNN dan menunjukkan peningkatan yang ketara dalam pengecaman emosi. Walau bagaimanapun, ketepatan keseluruhan masih agak rendah jika dibanding dengan kajian *Speech Emotion Recognition (SER)* sebelum ini. Cabaran utama termasuklah kesukaran dalam mengesan emosi disebabkan oleh ketidakseimbangan set data. Kerja masa depan mencadangkan kajian lanjutan memberi tumpuan kepada model Transformer, teknik penambahbaikan data yang lebih moden, dan kaedah pengesahan silang untuk meningkatkan generalisasi dan prestasi model.

Kata Kunci: Pengecaman Emosi Melalui Pertuturan, Pembelajaran Mendalam, Rangkaian Saraf Perlingkaran, Rangkaian Memori Jangka Pendek Panjang, CREMA-D, Pengekstrakan Ciri Suara, Pengesahan Silang.

Chapter 1 Introduction

1.1 Introduction

In today's technology-driven world, artificial intelligence has been very important in human daily life, bringing significant benefits and positive impacts across various industries. Currently, the needs of having AI technology like multimodal speech recognition is crucial in areas like healthcare, customer service and human computer-interaction. This kind of technology is very important because it enable system to interpret and respond to human communication more effectively by integrating multiple forms of input such as voice, text, and visual cues. This allows for a more direct and natural intuitive interaction between humans and machines, enhancing user experiences and satisfaction especially in those domains. A particularly advanced application of this technology is Speech Emotion Recognition (SER). Speech Emotion Recognition (SER) is the ability of a machine to understand and interpret human emotions (Singh et al., 2023). This involves analysing the tone, pitch, volume, and tempo of a person's voice to identify emotion state of individuals such as happiness, sadness, anger, or fear. Additionally, the process uses a variety of machine learning and signal processing techniques to extract meaningful patterns from the audio data of spoken words.

Emotion recognition from speech not only can enhance user experience, but at the same time also provide a useful insight especially in mental health applications and enable more empathetic AI systems. For example, in healthcare domain, recognizing emotions from speech can help in early detection of mental health issues, while in customer service, it can improve user satisfaction by tailoring responses based on emotional states.

However, although the need of this technology highly demand in current world settings, emotion recognition has remained a difficult challenge due to the difference on how individuals

express emotions through speech. Few challenges and factors such as cultural differences, linguistic diversity, voice accent and environmental noise making it difficult to develop a best model that generalize well across different contexts (Younis et al., 2024). Current emotion recognition systems often struggle to achieve high accuracy data and result due to these variations, which indirectly limiting their applicability in real-world applications.

This project is believed to address these challenges by utilizing advanced machine learning techniques. The CREMA-D dataset, comprising 7,442 labelled audio clips representing a wide range of emotions, serves as the foundation and testing dataset for this study (Cao, Cooper, et al., 2014). By leveraging deep learning architectures, this research aims to develop an acoustic model capable of recognizing emotions through spoken language.

This study is crucial as it contributes to the growing demand for AI systems that can interpret and respond to human emotions, which directly contributing to advancements in Human-Computer Interaction (HCI). By enabling systems to understand emotional tones in speech, the project has potential to enhances human-computer interactions, making AI more human-centric and intuitive. Additionally, this study also provides the opportunity of integrating Speech Emotion Recognition (SER) into Automatic Speech Recognition (ASR) to add a new dimension by not only transcribing spoken language into text but also analysing emotional nuances, paving the way for more effective applications in healthcare, customer service, and other domains.

1.2 Problem Statement

Although there are significant advancements in machine learning, the ability to accurately recognize emotions from speech remains a difficult challenge (Narimisaei et al., 2024). The variability in speech is influenced by few factors, including personal emotional expression, background noise, voice accent and linguistic diversity (Ding et al., 2012), (Tamati et al., 2013). These few factors often found to complicate the development of universal models that can effectively generalize across different language, speakers and contexts.

Besides, the primary challenge in developing SER systems is determining and choosing the suitable deep learning architecture. Current deep learning architectures such as Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and Transformers each architecture has unique capabilities to process and classify these emotional expressions from audio data (Y. Zhao & Shu, 2023). Thus, this project will analyse these architectures and their respective strengths and limitations in recognizing emotional expressions from speech, within the context of the speech emotion recognition.

Moreover, the existing datasets used for training and testing emotion recognition models often do not reflect the variety of emotional expressions found in real-world speech (Zhang & Pell, 2022). Emotions can be expressed differently across cultures. For example, what is considered a joyful expression in one culture may be interpreted differently in another (Hareli et al., 2015). Many current datasets do not include this cultural diversity, which can result in models that do not perform well across different populations (Costa et al., 2023). Therefore, this project aims to determine how well the CREMA-D dataset, supports the testing of acoustic models using suitable deep learning architecture.

1.3 Project Scope

This project focuses on advancing the field of emotion recognition from speech through the exploration and evaluation of deep learning architectures. The study will explore various deep learning architectures, including Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), Transformer models, and hybrid approaches, to evaluate their features and ability in recognizing emotions from speech. A comprehensive literature review will be conducted to compare the SER models developed and tested in the previous study. This comparison includes the deep learning architectures used in the SER model, the dataset employed to the study and the feature extraction techniques. These SER models will be analysed for their accuracy and reliability in recognizing emotional states based on audio input.

By referring to the comparison of previous study, a Speech Emotion Recognition Model will be developed based on recent trends of deep learning architecture and achieving accuracy of previous SER models. Next, this project will utilize the CREMA-D dataset, which consists of 7442 of labelled audio clips featuring a wide range of emotional expressions, including happiness, sadness, anger, and fear. This dataset will provide a rich resource for training and testing of the speech emotion recognition model that have been developed, ensuring that they are exposed to diverse emotional contexts and variations among speakers.

To ensure a comprehensive assessment of the model performance, the project will employ multiple evaluation metrics. These will include accuracy, precision, recall, and F1-score, providing insights into the strengths, accuracy and effectiveness of the model. By utilizing these metrics, the project aims to facilitate a detailed comparison of the effectiveness of various deep learning approaches in recognizing emotions from speech.

Lastly, an important aspect of this project is benchmarking the performance of the developed models on the CREMA-D dataset. This benchmarking process will yield valuable

insights into the applicability of different deep learning architectures for real-world emotion recognition tasks. By comparing model performance across various metrics, the project seeks to identify the most effective methods for accurately recognizing emotions in diverse contexts. Through these components, the project aims to significantly contribute to the understanding and advancement of emotion recognition technologies.

1.4 Research Questions

The research questions are as follows:

1) What are the recent trends in deep learning techniques applied in speech emotion recognition?
2) Which deep learning technique is the most suitable to be used for developing acoustic model and how this technique can recognize different emotions from CREMA-D dataset?
3) How does the performance of the trained model on the CREMA-D dataset compare with existing models?

1.5 Project Objectives

The specific objectives of this project are:

1) To analyse the deep learning technique (including CNNs, LSTMs, Transformer and hybrid approach) for emotion recognition.
2) To develop an acoustic model that able to recognize emotions (happiness, sadness, anger, etc.) from adult speech using the CREMA-D dataset.

3) To test the performance of the acoustic model using key metrics like accuracy, precision, recall, and F1-score, and benchmark them against existing studies.

1.6 Methodology

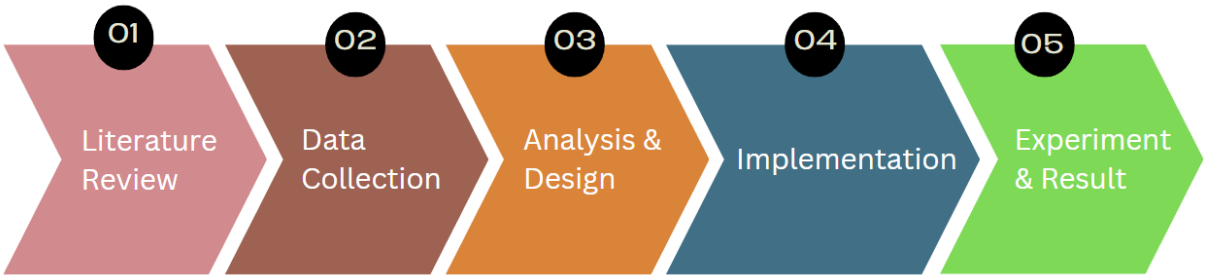


Figure 1: A structured methodology for the project

Figure 1 shows the methodology for the project titled "Emotion Recognition from Adult Speech Using Machine Learning: A Study Based on the CREMA-D Dataset". The methodology consists of five phases. Phase 1, Literature Review involves a thorough study and review of existing research on speech emotion recognition, the background, analysing various machine learning techniques and the extracting features like MFCC and Spectrograms. In Phase 2, Data Collection, the CREMA-D dataset is acquired, featuring 7,442 labelled audio clips. This data undergoes preprocessing to normalize and segment the audio suitable for model input. Next in Phase 3, Analysis and Design, follows choosing on the architecture to being used and designing for the deep learning model. This phase involves preparing the CREMA-D dataset, which is segmented into 70% for training, 10% for validation and 20% for testing with unseen data, to evaluate the model's performance. Implementation process is carried out in Phase 4, where an acoustic model is developed using a machine learning framework and using phyton language,

then testing on the dataset, and evaluation according to performance metric. Finally, Phase 5, Experimental and Results, focuses on evaluating the model using metrics such as accuracy, precision, recall, and F1-score to determine their effectiveness at emotion recognition from speech.

1.7 Significance of Project

The project on emotion recognition from adult speech using machine learning and the CREMA-D dataset aims to advance understanding and application in an important field of AI. By investigating various deep learning architectures and integrating insights from technology and psychology, this study contributes to both academic research and practical AI applications. It aims to deepen the understanding of how speech characteristics correspond to emotional states, establishing a foundational understanding for future advancements in AI systems.

1.8 Project Schedule

Task	Duration (days)	Start Date	End Date
Project Title, Overview and Discussion	9	02/10/2024	10/10/2024
Preparing and submitting Brief Proposal	6	10/10/2024	16/10/2024
Preparing and Submitting Full Proposal	27	17/10/2024	13/11/2024
Chapter 1: Introduction	39	13/10/2024	21/11/24

Chapter 2: Literature Review	53	21/10/2024	13/12/2024
Chapter 3: Requirement Analysis & Design	27	09/12/2024	05/01/2025
Amendment, compiling and submitting FYP 1 Report	12	05/01/2025	17/01/2025
Producing and submitting presentation video URL and presentation slides	5	13/01/2025	17/01/2025
Chapter 4: Implementation and Testing (Model Development)	55	10/3/2025	4/5/2025
Chapter 5: Result and Analysis (Model Evaluation)	11	5/5/2025	15/5/2025
Chapter 6: Conclusion and Future Work	16	16/5/2025	31/5/2025
Compiling and submitting FYP 2 Report	14	10/6/2025	23/6/2025
Compiling and submitting FYP 2 Report (After Amendment)	23	6/7/2025	28/7/2025

Table 1: Project Schedule of the research

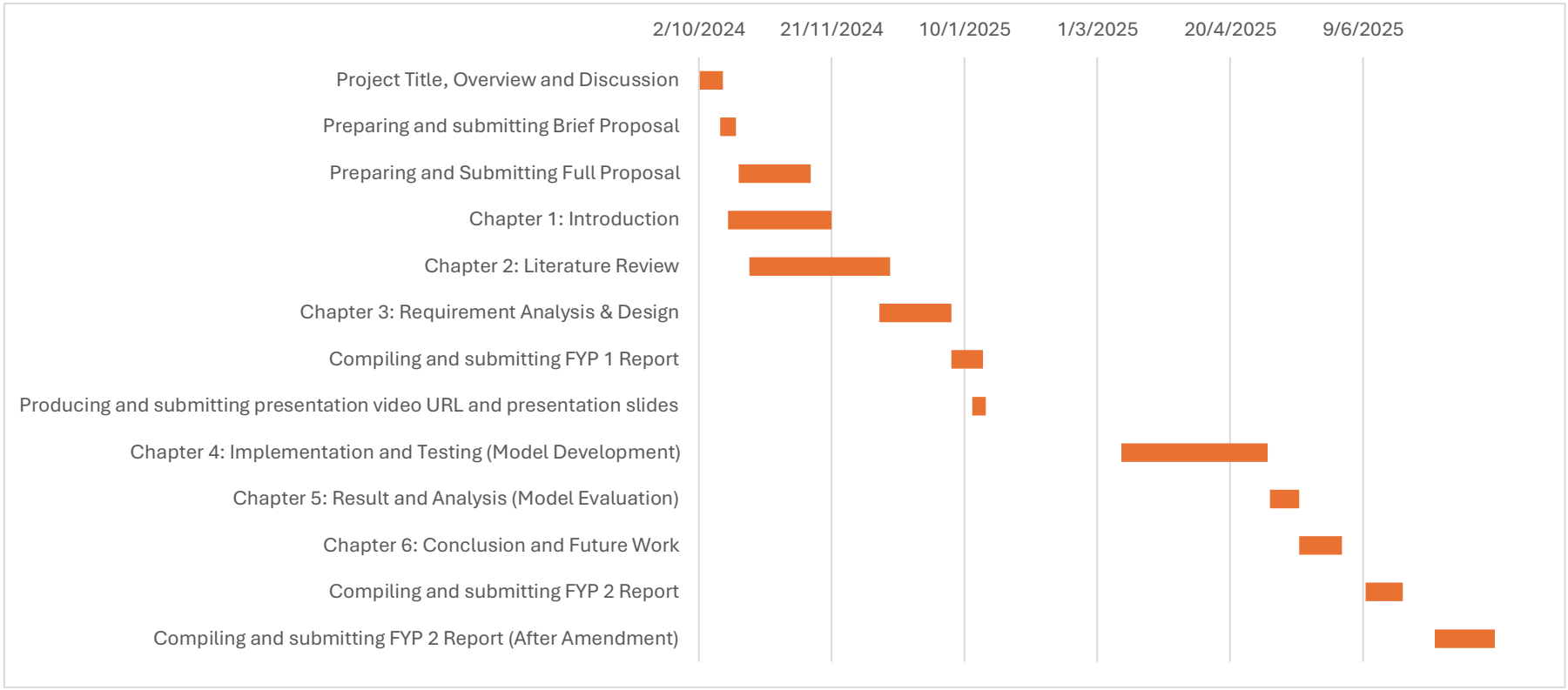


Figure 2: Gantt Chart of Project Schedule

1.9 Expected Outcome

The project is expected to deliver an acoustic model derived from one of the deep learning architectures that will be capable of recognizing emotions from the CREMA-D testing dataset. Along with a thorough assessment of deep learning models, it will provide practical insights and recommendations for further emotional identification studies. Additionally, the project aims to contribute to the development of more effective and practical applications for emotion recognition in areas such as customer service, mental health monitoring, and human-computer interaction.

1.10 Chapter Outline

The summary for next few chapters is as follows:

Chapter 2: Literature Review

In this chapter, a comprehensive review of existing research, theoretical foundation and methodologies related to speech emotion recognition is conducted. It includes review of previous studies that have utilized deep learning techniques (such as CNN, LTSM and Transformer), the common datasets that have been used in SER research, analysing feature extraction techniques (MFCC and Spectrogram) and basic tools for building an acoustic model. The chapter aims to establish a solid foundation for the proposed methodologies in subsequent chapters.

Chapter 3: Methodology

This chapter explains the methods that are being used in this project, the deep learning architecture that has been chosen and details the processes involved in data collection and

preprocessing from the CREMA-D dataset. Additionally, it will discuss the evaluation metrics employed to assess model performance.

Chapter 4: Implementation

This chapter covers the practical aspects of development and implementation of emotion recognition model based on the methodologies outlined previously. It presents the results of testing of the prototype (acoustic model), on the CREMA-D dataset, detailing their performance against established evaluation metrics such as accuracy, precision, recall, and F1-score. The chapter also includes a discussion on benchmarking these models against existing studies to evaluate their performance.

Chapter 5: Conclusion and Future Work

The concluding chapter summarizes the results and key findings from the research, validate the objectives of the project and discusses the implications of these findings for practical applications. Furthermore, potential future work is described, suggesting areas and opportunities for further research into enhancing model, exploring additional datasets and methodologies to improve emotion recognition capabilities.

1.11 Chapter Summary

This chapter employs a systematic framework commencing with an introduction to the project, continue with the problem statement, research questions, objectives, methodology, scope, significance of the project, expected outcomes, and a project overview. On the next chapter, it will include a thorough literature review, exploring related research to the project.

Chapter 2 Literature Review

2.1 Introduction

This chapter provides an overview of Speech Emotion Recognition (SER), discussing its evolutions, applications, and recent advancements of SER. This chapter also reviews the literature on various deep learning architectures, including Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, Transformers, and hybrid models, used for emotion recognition. Additionally, a comparative analysis of previous studies is also conducted to evaluate the accuracy of different SER models, the datasets used and feature extraction techniques. This also discuss about the evaluation metric for SER models, tools and frameworks used for implementing SER models.

2.2 Speech Emotion Recognition (SER)

2.2.1 Introduction to SER

Speech Emotion Recognition or SER refers to a process of identifying and interpreting human emotions from speech or any spoken language (G. Liu et al., 2023). It utilizes signal processing and various machine learning techniques to identify and classify emotional states such as happy, sad, angry, fear, surprise etc. This ability to identify emotional state through speech is particularly meaningful as emotion play a crucial role in human daily life, as it not only enable individuals to express their feelings and moods, but will also influencing human intelligence, decision-making, social interactions, perception, memory, and personal learning (G. Liu et al., 2023).

2.2.2 Applications of SER

SER has shown a significant impact in healthcare domain especially for mental health diagnosis and therapy. SER is being integrated into virtual assistants that support healthcare experts by identifying emotions such as sadness or anxiety through patients' voices (Guo et al., 2024; Singh et al., 2023). These systems analyse speech patterns to evaluate emotional states, which enable a quick treatment based on the specific requirements of the patients. (Singh et al., 2023). For example, during consultations, SER can monitor a patient's emotional changes, allowing healthcare experts to detect any early signs of mental health issues like depression or anxiety.

Next, SER is important in improving interactions between humans and computers or known as HCI (Alsabhan, 2023). By allowing systems to understand and react to users' emotions cues, SER can contribute to more engaging and intuitive interactions (Singla et al., 2024). For example, in educational setting, SER could help to adjust and align the study materials for students based on how they feel emotionally. This is important as it helps students

to increase their interest in study and improve memory retention. Research also shows that integrating emotional recognition into e-learning platforms can lead to better learning experiences by adjusting contents according to the emotional states detected from students' speech (Ma et al., 2023). This flexibility doesn't just improve users' emotions, but it promotes a more effective learning environment (Yap et al., 2021).

Besides that, SER also has been employed call centre analysis. By analysing customer interactions and emotional cues in conversations, SER could provide insights into how customers feel (Bojanić et al., 2020). This analysis helps in enhancing service quality by allowing call centre agents to align their responses based on caller's emotion state. Moreover, the integration of SER into customer service frameworks can improve customer satisfaction rates, as representatives become more adept at effectively addressing concerns (Almarzooqi, 2022; Chul Min Lee & Narayanan, 2005).

2.3 Foundations of Speech Emotion Recognition

2.3.1 Evolution of Speech Emotion Recognition

Emotions are fundamental to the human experience. From a psychological perspective, emotions can be understood as complex reactions that involve physical responses, thoughts about the situation, and behaviours to show how they feel (Izard, 2009). These emotional responses are important as they help in communication and social relations. One of the common ways on how human expressing their emotions is through speech.

In the early study of emotion recognition through speech, the focus was primarily on identifying the acoustic cues and patterns in speech that are associated with different emotions (Scherer, 2003). This study is to explore the relations between acoustic features such as pitch, intensity, and tempo, and the emotional states conveyed in the speech (Ekberg et al., 2023; Scherer, 2003). As technologies continue to advance, affective computing began to develop as a field that combines various disciplines including computer science, psychology, and linguistics, which aim to help computers to recognize human's emotions (Pei et al., 2024).

In the 2000s, the advancement of machine learning has led to the application of statistical models such as Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) for speech emotion recognition (Nogueiras et al., 2001). Researchers start to explore and use these machine learning techniques to capture the temporal dynamics and probabilistic relationships between acoustic features and emotions (Schuller et al., 2003). These models classify emotion states based on the learned statistical patterns.

2.3.2 Basic Workflow of SER Systems

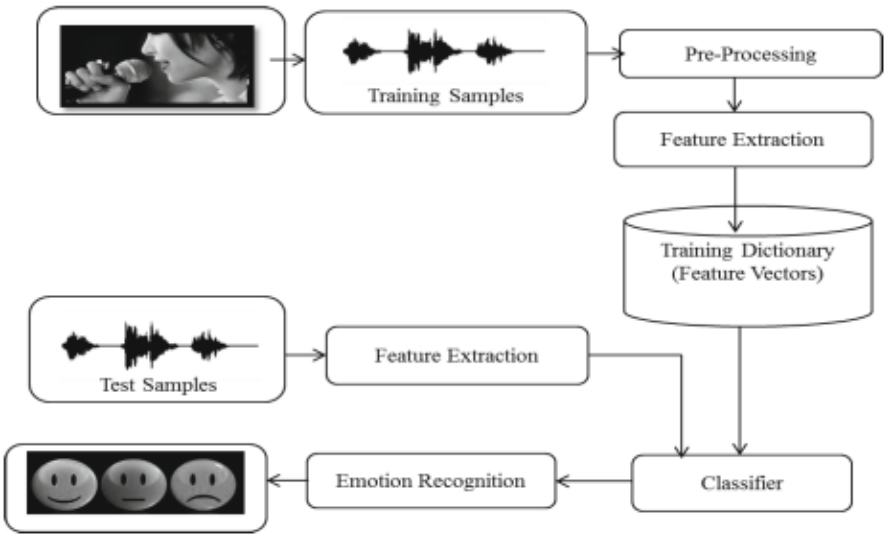


Figure 3: Workflow of SER system using machine learning technique (Kamble et al., 2015).

Speech Emotion Recognition (SER) system usually consist of five main components which includes data gathering and recording, pre-processing to improve signal quality, feature extraction, classification and emotions recognition. The architecture of speech emotion recognition system is as shown in Figure 1. The process begins with collection of audio recordings referred to as training samples. These training samples contain speech with emotional expressions, usually categorised with relevant emotional categories.

Generally, speech emotion recognition involves a complex interplay of acoustic features and machine learning techniques (Rezapour Mashhadi & Osei-Bonsu, 2023). First, audio recordings are pre-processed to enhance signal quality and reduce noise (Hasija et al., 2022; Ibrahim et al., 2019). Next, to extract feature vectors from the input speech, algorithms such as Mel-frequency Cepstral Coefficients (MFCC) and prosodic are usually used to help to extract and analyse the spectral and temporal characteristics of speech, which enabling the accurate identification of emotional states through variations in tone, pitch, rhythm, and intensity (Patil

et al., 2012; Tracey et al., 2023). Finally, there will be emotion classification and recognition phase using machine learning technique such as Gaussian mixture Model (GMM) and Support Vector Machines (SVM) (Patil et al., 2012). The technique chosen for the system is based on the suitability of the model to the unique attributes and complexity of the dataset (X. Liu et al., 2023).

However, the accuracy of the SER system is relied on the naturalness of the dataset which is used as an input to the system. The dataset as an input to the system may contain the real-world emotions or the acted ones. Thus, it is more practical to use dataset that is collected from the real-life situations (El Ayadi et al., 2011). The commonly used SER datasets are shown in section 2.6 (Overview of datasets for SER tasks).

2.3.3 Recent Advances in SER

Recent trends in Speech Emotion Recognition (SER) show significant advancements especially with the integration of deep learning techniques (Khalil et al., 2019). Deep learning architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LTSM) has demonstrated high effectiveness in extracting emotional cues from speech (Islam et al., 2024). For example, by utilizing CNNs, it could automatically extract distinctive features from speech spectrograms, which enable models to achieve higher accuracy in classifying emotions. These models have outperformed methods such as Gaussian mixture Model (GMM) and Support Vector Machines (SVM), which has achieving higher accuracy rates in detecting basic emotions such as anger, joy, and sadness (Islam et al., 2024). This evolution has led to improved accuracy in recognizing emotional states, making SER become more reliable and efficient in various applications.

2.4 Deep Learning Architectures in Speech Emotion Recognition

2.4.1 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) is a deep learning architecture that is designed to process structured data like images and audio signals (Chu et al., 2023). One of the key components in CNNs is its convolutional layers which is used to automatically extract features from the input data and allowing them to recognize the patterns. This capability makes CNNs particularly effective for tasks such as image classification, object detection, and speech recognition (Chu et al., 2023; Wolf-Monheim, 2024). In speech emotion recognition, CNNs are used to analyse audio signals by converting them into spectrograms, which are visual representations of sound frequencies over time (Nanni et al., 2021). By applying convolutional layers, the network learns to identify emotional cues in speech, such as variations in tone and pitch, which helps accurately classify the emotions conveyed in spoken language. This ability to learn complex patterns from raw data highlights the strength of CNNs in both visual and auditory applications.

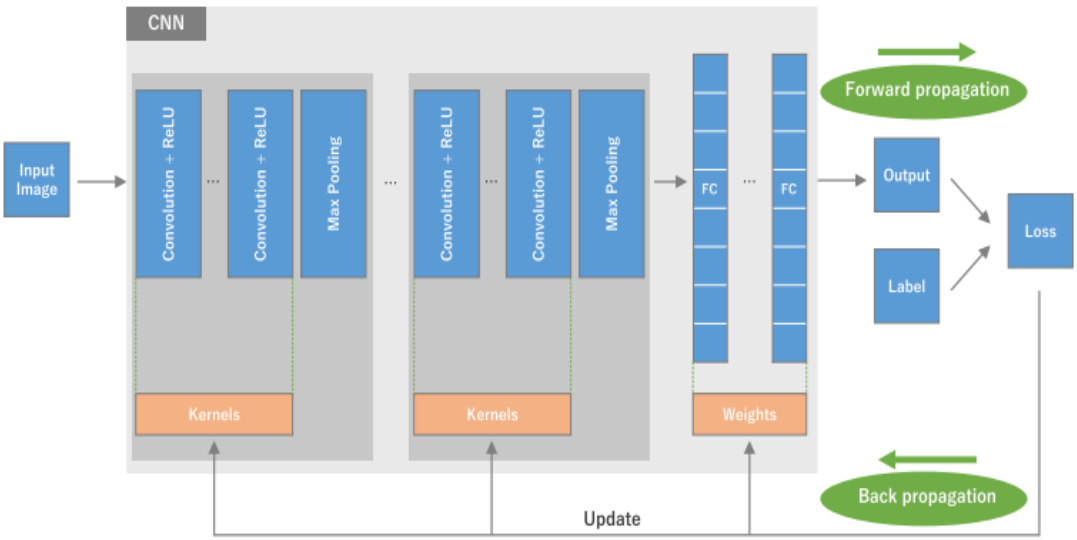


Figure 4: An overview of a convolutional neural network (CNN) architecture (Yamashita et al., 2018).

CNNs consist of several key components that work together to extract features from input data. This architecture has four main components which includes convolutional layers, Activation Function, pooling layers, and fully connected layers as shown in Figure 2. The first component is the convolutional layer, which serves as the backbone of a CNN and is responsible for feature extraction (Nanni et al., 2021). It applies a set of filters (or kernels) to the input data before the data been used. By stacking multiple convolutional layers, CNNs can learn increasingly complex features, starting from simple edges in early layers to more complex shapes and objects in deeper layers.

After each convolution operation, an activation function is applied to introduce non-linearity into the model. This step allows the network to learn complex patterns in the data. The commonly used activation function in CNNs is called the Rectified Linear Unit, or ReLU. ReLU is a mathematical function used in deep learning models to introduce non-linearity in neural networks (Bai, 2022). It works simply by transforming any negative values to zero while keeping positive values unchanged (Yamashita et al., 2018). ReLU helps speed up the training process and enables the network to learn complex relations within the data (Bai, 2022).

$$f(x) = \max(0, x)$$

Equation 1: Mathematical function of ReLU (Bai, 2022).

Next component will be Pooling layers. Its main function is to perform downsampling or sub-sampling on the feature map generated from the previous convolution operation. Their main function is to reduce the spatial dimensions (width and height) of the feature maps while preserving important information, which decreases the amount of computation needed and helps prevent overfitting (Zhou et al., 2023). By summarizing features within specific regions

of the feature map, pooling layer converts a large feature map into a smaller one while retaining important information or features.

The fully connected layer (FC layer) comes after several convolutional and pooling layers and serves as the final stage of feature extraction and categorization (Goh et al., 2024). Every single neuron in this layer is linked to every single neuron in the previous layer, allowing it to process the spatial data that has been acquired from the earlier layers to the intended output classes (Srinivas et al., 2024). FC layers facilitate complex pattern recognition and decision-making, enabling CNNs to perform tasks like image classification, object detection, and speech emotion recognition (Srinivas et al., 2024).

2.4.2 Long Short-Term Memory (LSTM) Networks

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) designed to handle sequences of data, making them particularly effective for tasks involving time-dependent information, like speech or text. LSTMs can remember important information over long periods of time while forgetting irrelevant data, which helps them understand context in sequences, such as sentences or audio waves (Sak et al., 2014). In deep learning, LSTMs are used to process and predict time-series data as traditional RNNs have limitation with long-term dependencies (Sak et al., 2014). In speech emotion recognition, LSTMs are useful because they can capture the temporal aspects of speech, like changes in pitch, tone, and rhythm, which are crucial for detecting emotions such as joy, sadness, or anger in a person's voice. By analysing these speech patterns over time, LSTMs can effectively identify the emotional state behind the spoken words (Kurniawan et al., 2023).

One of key features in LSTM networks is their use of cell states, which act as a form of memory (Malashin et al., 2024). The cell state is like a conveyor belt that carries information

throughout the network. It runs through the entire chain of LSTM units and allowing data to flow from one time step to the next. This design is crucial because it helps the network remember important information and details over a long period.

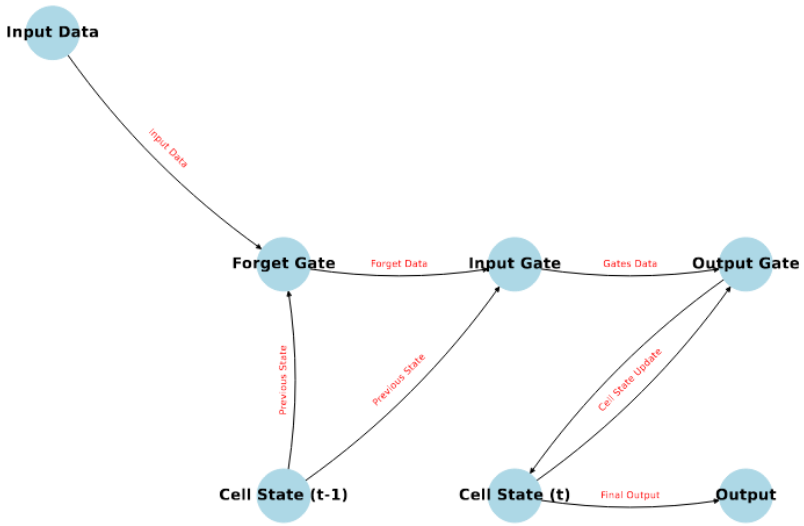


Figure 5: The Architecture Diagram of LSTM (Malashin et al., 2024).

The architecture of LSTMs also includes a unique gating mechanism that consists of three types of gates: the input gate, the forget gate, and the output gate as shown in Figure 3. These gates control how information flows into and out of the memory cell. The input gate decides how much of the new information coming into the LSTM should be added to its memory by using a mathematical function called the sigmoid activation function as shown below, where σ is the sigmoid function, Wi represents the weight matrix, ht −1 is the previous hidden state, xt is the current input, and bi is the bias (Lu & Salem, 2017; Malashin et al., 2024).

$$it = \sigma(Wi \cdot [ht - 1, xt] + bi)$$

Equation 2: Mathematical Equation of Sigmoid Activation Function (Malashin et al., 2024).

On the other hand, the forget gate decides what information from the past should be discarded. This helps the LSTM remove outdated or irrelevant information, ensuring that only useful data influences future predictions (Lu & Salem, 2017). Lastly, the output gate controls what information from the memory cell will be sent out as output at each time step. It decides which parts of the stored data are relevant for making predictions or classifications based on current inputs.

2.4.3 Transformer

Transformers are an advance type of neural network architecture designed to work with sequences of data and effective for tasks like natural language processing and speech emotion recognition. Introduced in the paper "Attention is All You Need" by Vaswani et al. (2017), Transformers have changed the perspective of deep learning by eliminating the need for recurrent structures that previously were essential for processing sequences. Instead, this architecture rely entirely on a self-attention mechanism which allowing them to capture relationships between all elements in a sequence simultaneously, regardless of their distance from one another (Vaswani et al., 2017).

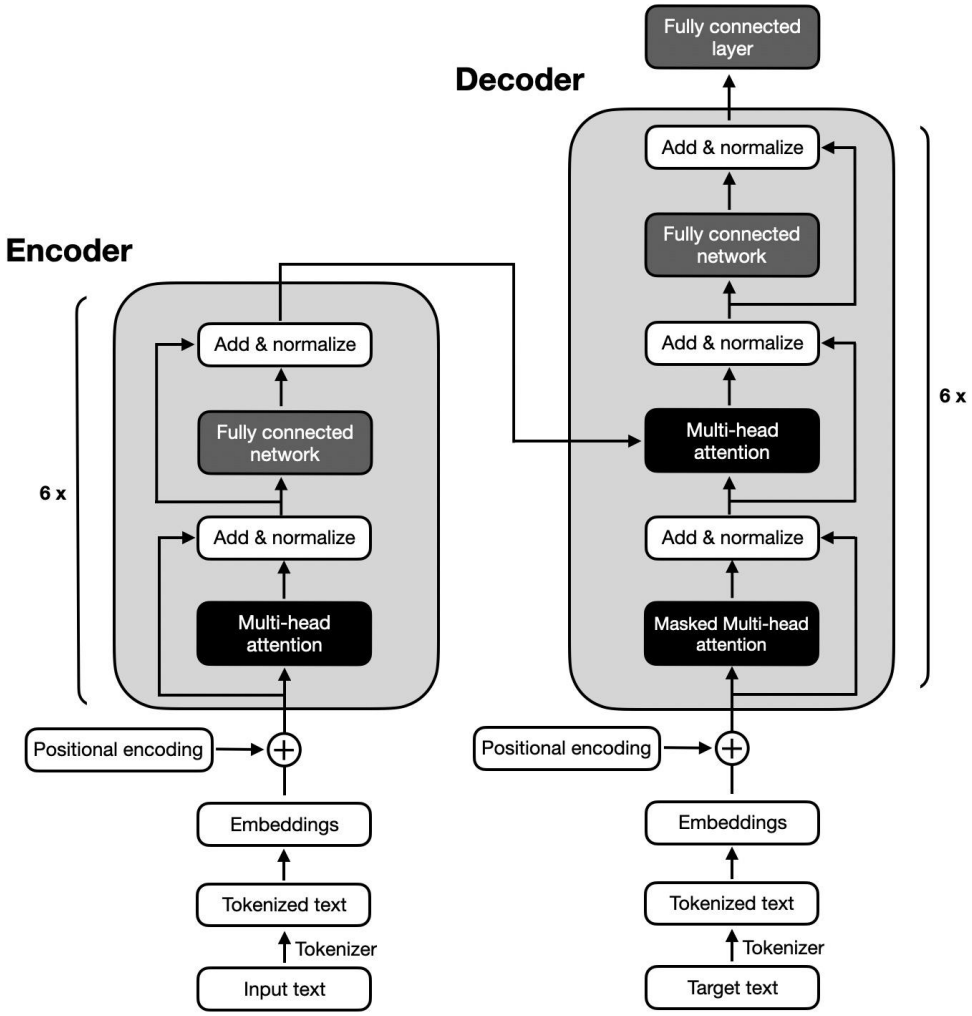


Figure 6: The architecture of transformer model (Vaswani et al., 2017).

Self-attention mechanism plays an important role in Transformer as it enables the model to determine how important different parts of the input data relative to each other (Ambartsoumian & Popowich, 2018). This mechanism helps the model to create a comprehensive understanding of the input data and capturing relationships between distant elements. This mechanism is useful for tasks that involve sequential data, such as speech recognition or translating languages, as it can more effective capture relationships between

distant elements in the sequence compared to models like RNNs or LSTMs (Ambartsoumian & Popowich, 2018).

One of key components in Transformer is the Encoder-Decoder architecture as shown in Figure 3. The encoder's role is to process the input data and transforms it into continuous representations that reflect the context of the input (Vaswani et al., 2017). The encoder consists of multiple layers, with each containing a multi-head self-attention mechanism and a feed-forward neural network, which work together to refine these representations (Sonkar & Baraniuk, 2023). The encoder's final output is a set of continuous representations, which are then fed into the decoder. The decoder then generates the model's output from these representations.

Besides, Multi-Head Attention in Transformer improve the model's capability to capture complex relationships within the data (Vaswani et al., 2017; Voita et al., 2019). Instead of focusing on just one aspect of the input at a time, multi-head attention splits this attention into several parallel "heads.". Each attention head operates on a linear projection of the queries, keys, and values, and their outputs are concatenated and transformed into a final output. This parallelization allows the model to attend to various parts of the input simultaneously, improving its ability to understand the diverse dependencies present in the data.

Another key aspect of the Transformer model is Positional Encoding. Since Transformers process tokens in parallel and do not inherently capture the order of tokens in a sequence, positional encodings are added to the input embeddings to provide information about the relative or absolute position of each token (Chen et al., 2021; Sajun et al., 2024). The positional encodings are designed using sinusoidal functions of different frequencies, allowing

the model to easily differentiate token positions and enabling it to extrapolate to longer sequences during inference.

In each layer of a Transformer, the output from self-attention mechanisms is processed by Feed-Forward Neural Networks. These networks use two linear transformations with a ReLU activation function in between. This process helps capture complex relationships and improves the model's understanding of the input data (Vaswani et al., 2017).

2.4.4 Hybrid Model

Hybrid model in deep learning architecture often refers to the combination of different machine learning techniques. The combination of different techniques is usually to optimise the strength of each architecture and improve the overall performance of the model.

For instance, the combination of the architecture of Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs) as a hybrid model (CNN-LSTM) is often being used to process sequential data that has both spatial and temporal characteristics (Abdallah et al., 2021). Basically in these models, CNNs are utilized to extract spatial features from input data, such as images or audio signals, by applying convolutional layers that identify patterns and structures (Nanni et al., 2021). Once the CNN has extracted relevant features, these are fed into LSTM layers, which are designed to capture temporal dependencies in the data (Sak et al., 2014).

This two-step process in CNN-LSTM architectures is particularly suited for applications such as visual time series prediction or generating textual descriptions from sequences of images (Abdallah et al., 2021). By integrating CNNs and LSTMs in an end-to-end architecture, these hybrid models can learn to recognize complex patterns in data, at the same time

understanding how those patterns evolve over time. This combination allows for a more comprehensive analysis of data compared to using either architecture alone. For example, in speech emotion recognition, the combination of these architectures is particularly useful as CNNs can identify key features from audio spectrograms, while the LSTM can interpret how these features change over time to accurately classify emotions (Salian et al., 2021).

2.4.5 Comparison of previous study

This section summarizes the previous study of SER in a table form. The table shows a comparison of the previous study based on the deep learning architectures used, features extraction techniques used, the dataset used for each of the study and the results obtained.

Research Paper	Architecture Used	Features Used	Dataset	Result (Accuracy (%))
Learning deep features to recognise speech emotion using merged deep CNN (J. Zhao et al., 2018)	CNN	Log-Mel Spectrogram	IEMOCAP, EMO-DB	85.66, 88.22
Speech Emotion Recognition Using 1D CNN with No Attention (Li et al., 2019)	CNN	MFCC, Mel Spectrogram,	EMO-DB, RAVDESS, IEMOCAP	86.67, 75.25, 65.35
Performance Comparison of LSTM Models for SER (Swain et al., 2021)	LSTM	Spectral features, Pitch features	RAVDESS	76.83
Robust speech emotion recognition	CNN-LSTM	Log-Mel Spectrogram	RAVDESS, IEMOCAP,	99.47 (RAVDESS), 98.13

using CNN + LSTM based on stochastic fractal search optimization algorithm (Abdelhamid et al., 2022)			EMO-DB, SAVEE	(IEMOCAP), 99.76 (EMO-DB), 99.50 (SAVEE)
Performance Improvement of Speech Emotion Recognition Systems by Combining 1D CNN and LSTM with Data Augmentation (Pan & Wu, 2023)	CNN-LSTM (With Data Augmentation)	MFCC	RAVDESS, EMO-DB, IEMOCAP	95.52, 95.84, 96.21
Wavelet Multiresolution Analysis Based Speech Emotion Recognition System Using 1D CNN LSTM Networks (Dutt & Gader, 2023)	CNN-LSTM (With Wavelet Analysis)	Wavelet-based feature extraction	RAVDESS	81.2 (Weighted Accuracy), 81.4 (Unweighted Accuracy)
Speech emotion recognition using deep 1D & 2D CNN LSTM networks (J. Zhao et al., 2019)	CNN-LSTM	Log-Mel Spectrogram	IEMOCAP, EMO-DB	1D CNN LSTM-67.92 (IEMOCAP), 92.34 (EMO-DB). 2D CNN LSTM-89.16 (IEMOCAP), 95.33 (EMO-DB).
CREMA-D: Improving Accuracy with BPSO-Based Feature Selection for	CNN-BPSO-SVM (BPSO stands for Binary Particle Swarm Optimization)	MFCC	CREMA-D	66.01

Emotion Recognition Using Speech (Donuk, 2022).				

Table 2: The Review and Comparison of Previous Study

From the Table 1 above, most of the studies have utilizing the hybrid deep learning models which combining CNNs and LSTMs. This hybrid model generally achieves good performance in speech emotion recognition as compared to individually used architecture, with this hybrid model can achieve accuracy ranging from 69% to almost 99%. Furthermore, feature extraction techniques such as Log-Mel spectrograms and MFCCs seems to be particularly effective in achieving high accuracy, especially in datasets such as IEMOCAP and Emo-DB.

2.5 Common Feature Extraction for Emotion Recognition

Feature extraction refers to process of identifying and extracting voice characteristic from audio signal to classify emotion (Labied & Belangour, 2021). Mel-Frequency Cepstral Coefficients (MFCCs) are a common and widely used feature extraction technique in speech and audio signal processing, particularly in applications like automatic speech recognition, speaker identification and emotion recognition (De Lope & Graña, 2023). MFCCs are designed to represent the short-term power spectrum of sound based on a nonlinear Mel scale, which closely aligns with the way humans perceive sound frequencies (Rezapour Mashhadi & Osei-Bonsu, 2023). MFCCs are known for its ability to mimic human auditory system and capture the unique qualities (timbre) and sound patterns (harmonic) in speech.

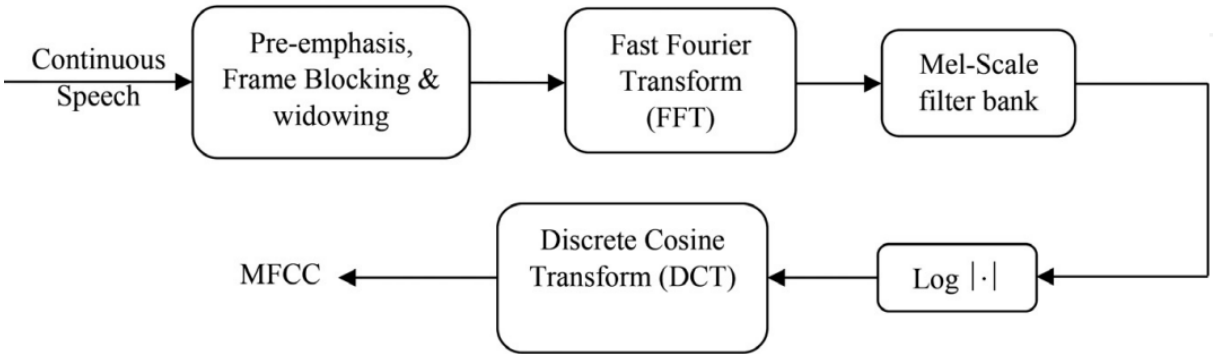


Figure 7: The block diagram of MFCCs (Ajibola Alim & Khair Alang Rashid, 2018).

Generally, MFCCs computation involves several steps and processes to obtain the needed coefficients as shown in Figure 5. MFCCs are effective in capturing vocal tract resonances in the low-frequency range, making them relevance in identifying emotional content in speech (Ajibola Alim & Khair Alang Rashid, 2018).

Prosodic features refer to the vocal characteristics that carry information about the speaker's emotions through variations in pitch, intensity, tempo and duration of the speech (Mary & Yegnanarayana, 2008). Pitch refers to perceived frequency of the voice or how high or low the voice sounds. Higher in pitch often indicates someone is excited. Intensity refers to the loudness or volume of the sound. It also represents the energy level of the speech signal. Higher speech volume often reflects to emotion of anger. Duration refers to the length of speech segments and can indicate how certain sounds are emphasized or stretched. By analysing all these features, it helps to identify emotion states more accurately as prosodic allows the model to capture the nuances of how emotions are expressed through intonation, rhythm and timing (Cao, Beňuš, et al., 2014).

2.6 Overview of Speech Emotion dataset

This section summarizes the commonly used SER datasets in a table form. The table shows a comparison of the datasets including the number of samples, number of speakers, languages for the dataset, emotions covered and the source accessibility.

Dataset	Description	Number of Sample	Number of Speakers	Language	Emotions covered	Source Accessibility
CREMA-D (Cao, Cooper, et al., 2014)	The clips were collected from 48 male and 43 female actors between the age of 20 and 74 across a variety of races and ethnicities.	7442	91	English	Anger, Disgust, Fear, Happy, Neutral, Sad	Open Source
RAVDESS (Livingstone & Russo, 2018)	A gender-balanced dataset that consists of 24 professional actors in a neutral North American accent	1440	24	English	Calm, Happy, Sad, Angry, Fearful, Disgust, Surprised, Neutral.	Open source
TESS (Pichora-Fuller & Dupuis, 2020)	A collection of audio samples from two female actors aged 26 and 64 years old	2800	2	English	Anger, Disgust, Fear, Happiness, Pleasant Surprise, Sadness, Neutral	Open source
EmoDB (Burkhardt et al., 2005)	A speech dataset that contains audio recordings of 10 professional German	535	10	German	Anger, Boredom, Disgust, Fear, Happiness, Sadness, Neutral	Open Source

	actors (5 male and 5 female)					
SAVEE (Jackson & Haq, 2011)	A dataset of audio samples from four male actors aged between 27 and 31 years old	480	4	English	Anger, Disgust, Fear, Happiness, Sadness, Surprise, Neutral	Open source

Table 3: The Comparison of Common Speech Emotion Datasets

From the comparison of dataset in Table 2, CREMA-D dataset has the largest audio samples with 7442 clips and containing recordings from 91 diverse speakers (48 male and 43 female), which offering variety in voice characteristics, accents, and speaking styles. Additionally, CREMA-D also covers 6 different basic emotions including Anger, Disgust, Fear, Happy, Neutral, and Sad. Lastly, this dataset is suitable for testing acoustic model because language being in the recordings is English, which is a widely used and common language in daily life and easy to understand. Thus, CREMA-D dataset seems to be the most suitable for testing a speech emotion recognition system (acoustic model).

2.7 Evaluation Metrics in SER

In Speech Emotion Recognition, the commonly used performance metrics to evaluate the performance of the acoustic model are includes accuracy, precision, recall and F1.

Each of these metrices are calculated as follows:

1. Precision

   Precision refers to the ratio of correctly predicted positive observations to the total predicted positives. Precision can be calculated using formula below:

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

Equation 3: Formula to calculate Precision (Hu & Thing, 2024).

2. Recall

   Recall refers to the ratio of correctly predicted positive observations to all observations in the actual class. Recall can be calculated using formula below:

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

Equation 4: Formula to calculate Recall (Hu & Thing, 2024)

3. Accuracy

   Accuracy is a measure of the overall correctness of the model. It is the ratio of correctly predicted instances to the total instances. Accuracy can be calculated using formula below:

$$Accuracy = \frac{TruePositive + TruePositive}{TruePositive + FalsePositive + TrueNegavtive + FalseNegative}$$

Equation 5: Formula to calculate Accuracy (Hu & Thing, 2024)

4. F1 Score

   F1 score is the harmonic mean of precision and recall. It provides a balance between precision and recall. F1 score can be calculated using formula below:

$$F1\ Score = 2 * \frac{Precision * Recall}{Precisison + Recall}$$

Equation 6: Formula to calculate F1 Score (Hu & Thing, 2024)

2.8 Tools for experiments

2.8.1 Programming Language

Python is a popular programming language widely used in artificial intelligence (AI) and deep learning because of its simplicity, readability and versatility. It also provides a wide range of framework and libraries such as TensorFlow and Keras that make it ideal for developing complex models efficiently.

2.8.2 IDE Software

Visual Studio Code is a lightweight IDE that supports multiple programming languages, including Python. It offers features like debugging, code completion, and extensions, which make it an excellent choice for developing and testing deep learning models.

2.8.3 Deep Learning Framework/Library

TensorFlow is an open-source framework developed by Google for machine learning and deep learning applications. It allows developers to build and train models that can analyse large datasets and perform tasks like image recognition, natural language processing, and predictive analytics. Moreover, TensorFlow is also particularly useful for building deep learning models that can analyse audio signals and detect emotions. It provides the necessary tools to build

neural network architectures and preprocess audio data. Keras is a high-level neural networks API that runs on top of TensorFlow. It simplifies the process of building and training deep learning models by providing an intuitive interface and pre-built functions.

2.9 Motivation of research

Artificial intelligence has become very common in human daily life. It has been integrated in various domains including manufacturing, education and entertainment to improve efficiency, user experiences and enable automation of complex task. In healthcare and customer service, the ability to recognize and interpret human emotions from speech has become very essential. However, emotion recognition systems often facing challenges due to variations in emotional expression influenced by factors such as different languages and personal emotional expressions, which limit their effectiveness in real-world applications. Using CREMA-D dataset, this study aims to address on these gaps by utilizing suitable deep learning architectures such as CNNs and LSTMs and testing the techniques to build acoustic model for classifying human emotions. Besides, curiosity and personal interest in exploring AI applications has also become motivation for us to explore deeper and more on this topic specifically how machine learning could identify and classify human emotions from speech. This study is expected enhance the understanding of emotion recognition technologies.

2.10 Summary

This chapter consists of summary of literature review which related to the general overview of SER, its applications and the evolution of SER. It also consists of review of deep learning architectures, features extraction techniques and common datasets used in speech emotion recognition.

Chapter 3 Methodology

3.1 Introduction

This chapter discusses and explain about the methodologies used in this research, including data collection, data transformation, data augmentation, feature extraction techniques, the proposed speech emotion recognition model, and model evaluation as shown in Figure 6 below.

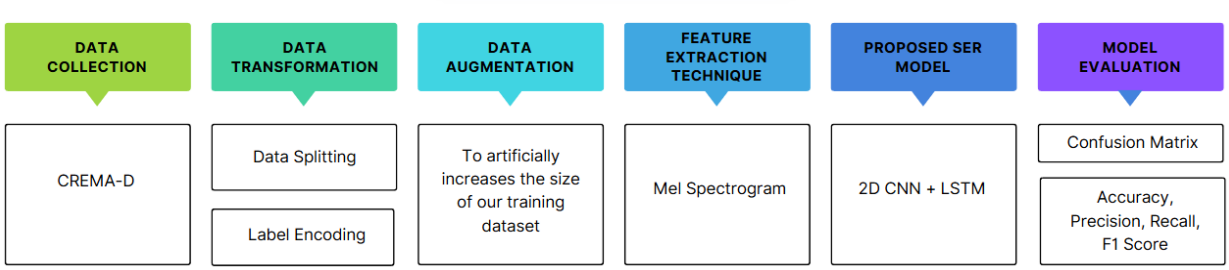


Figure 8: The Methodology of the research

3.2 Data Collection

This research has utilizing CREMA-D, one of widely used speech emotion datasets for models training and testing. This dataset is open-source dataset and can be download from kaggle.com. CREMA-D contains 7442 audio clips recorded from total number of 91 actors. These clips were from 48 male and 43 female actors between the ages of 20 and 74 coming from a variety of races and ethnicities (Cao, Cooper, et al., 2014). These actors spoke from a selection of 12 sentences. The sentences were presented using one of six different emotions including anger, disgust, fear, happy, neutral and sad.

3.3 Data Transformation

3.3.1 Data Splitting

In this research, the dataset is split into training and testing subsets. 70% of the data from CREMA-D are used for training. This training set is important in helping the model to learn to recognize different emotions by analysing patterns within this data. The remaining 10% and 20% of the data is used for validation and testing respectively. This testing set is used to evaluate the model's performance and assessing how well the model can generalize on new, unseen data. This data splitting ensure that the model learns to recognize emotions effectively rather than simply memorizing the training examples.

3.3.2 Label Encoding

Label Encoding is a process of converting categorical emotion labels, such as "happiness," "sadness," and "anger," into numerical values that can be used by machine learning models (Stanley et al., 2023). This transformation is very useful as most machine learning algorithms require numerical input to perform calculations and learn from data. Thus, by assigning a unique integer to each label, we can create a format that can be interpreted easily by the model. For example, emotions like "Happiness" may be encoded as 0, "Sadness" as 1, and "Anger" as 2. This numerical representation allows the model to recognize and differentiate between various emotions based on their encoded values.

3.4 Data Augmentation

Deep learning models often require a large amount of training data to perform effectively, but gathering such extensive datasets can be difficult. To overcome this issue, we use a technique called data augmentation. Data augmentation is a technique to artificially increases the size of

training dataset by creating modified versions of the original audio samples (Atmaja & Sasou, 2022). This modifications include things like add background noise to mimic real-world conditions, slightly change the pitch of an audio sample while maintaining its emotional content or adjust the playback speed without altering the pitch (Pan & Wu, 2023). This not only increases the dataset size but also helps to improve the generalization ability of the model.

3.5 Feature Extraction Technique

Mel spectrogram is a time-frequency representation of an audio signal that transforms the raw audio into a visual spectrum, capturing both the temporal and spectral features of speech (Sharan et al., 2024). It applies the Mel scale, which mimics the way humans perceive sound and emphasizing lower frequencies that are more relevant for speech and emotion recognition. In Speech Emotion Recognition models, Mel spectrograms are used as features because its ability to capture key prosodic elements such as pitch, energy, and speech rate. By converting the amplitude of frequencies into a logarithmic scale and using a Mel filter bank, the Mel spectrogram effectively highlights the emotional nuances in speech, making it as a powerful feature for emotion recognition tasks.

3.6 Proposed Speech Emotion Recognition Model

3.6.1 Architecture Design

Previous studies on Speech Emotion Recognition (SER) models have shown that combining CNN and LSTM architectures significantly enhances accuracy in emotion recognition compared to using either architecture independently. For example, research by Abdelhamid et al. (2022) and Zhao et al. (2019) using hybrid model of CNN and LSTM have achieved an

amazing accuracy with more than 80% of emotion recognition on Emo-DB and IEMOCAP dataset.

Thus, building on these encouraging findings, we propose a hybrid SER model that integrates CNN and LSTM architectures for this research. This proposed model will be tested on the CREMA-D, a dataset that contains a wide variety of emotional speech samples with total 7442 audio clips, ensuring a thorough evaluation of the model's performance.

3.6.2 Model description

Specifically, the proposed model combines a 2D CNN architecture, an LSTM architecture, and fully connected layers, as shown in Figure 6 below. A 2D CNN is a deep learning architecture designed to process two-dimensional data, making it particularly suitable for spectrograms, which represent audio data in time and frequency dimensions (Begazo et al., 2024). This allows 2D CNNs to efficiently extract features through convolution operations. Additionally, for large datasets like CREMA-D, containing 7,442 audio clips, 2D CNNs are highly effective due to their ability to process large amounts of data in parallel, generalize to diverse patterns, and reduce overfitting while maintaining computational efficiency.

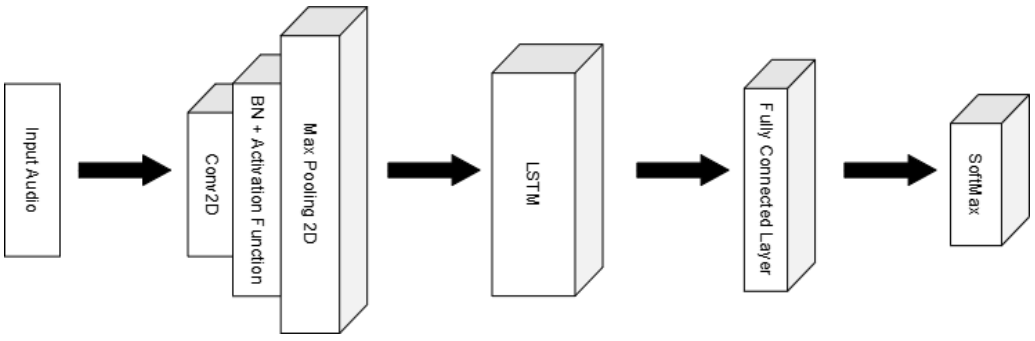


Figure 9: The design of proposed SER model.

The process begins with audio preprocessing, where audio files are transformed into spectrogram representations. These spectrograms capture the time-frequency characteristics of sound, providing a visual depiction of how the audio signal varies over time.

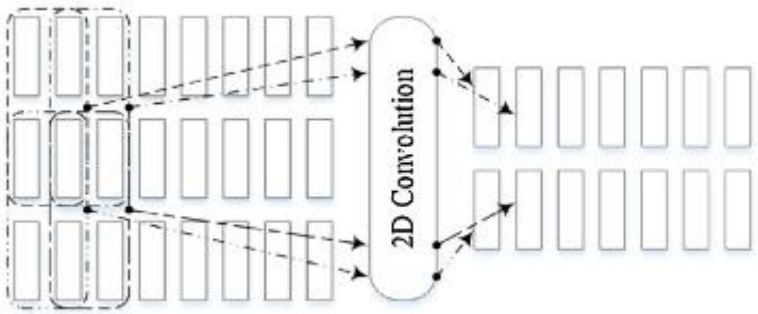


Figure 10: The Illustration of 2D convolution (J. Zhao et al., 2019).

The first component, Conv2D layer, is responsible for applying convolution operations to extract both spatial and temporal features from the spectrogram of the input audio signal. In this process, a set of learnable filters, also known as kernels, slides over the 2D input, extracting local time-frequency patterns that are required for recognizing emotions (Ayadi, 2024). These patterns may include variations in pitch, intonation, and energy levels over time. Each filter generates a feature map by performing a convolution operation, which involves calculating the dot product of the filter weights and patches of the input data (Ayadi, 2024). To enhance the performance, batch normalization and activation functions are applied to the model. Batch normalization normalizes the inputs of each layer. This is to helps maintain consistent feature distributions throughout training, leading to faster convergence and improved model's stability (Sefara, 2019). On the other hand, activation functions introduce non-linearity into the model, allowing it to learn complex patterns in the data.

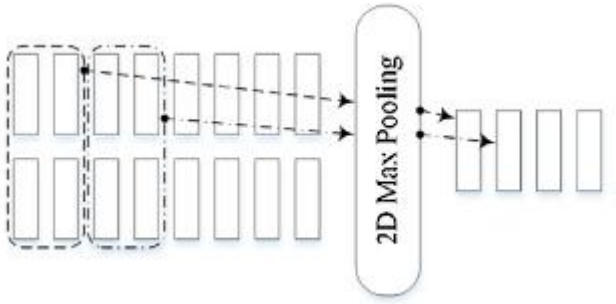


Figure 11: The Illustration of 2D Max Pooling (J. Zhao et al., 2019).

Following this, 2D max pooling is performed, which reduces the spatial dimensions of the feature maps while preserving the most significant features. Max pooling makes the model more computationally efficient and reduces overfitting by making the model less sensitive to small shifts in the input data (Sharma & Mehra, 2019). During the pooling operation, a sliding window moves across the input feature maps, selecting the maximum value from each window. This selection process effectively down-samples the feature maps, decreasing their height and width while keeping the number of channels constant (Sharma & Mehra, 2019). The result is a set of reduced feature maps that emphasize the most prominent patterns which are then passed to subsequent layers (LSTM) for sequential processing.

The output from the CNN block is then passed to the LSTM block to capture temporal dependencies and sequence information that are important for understanding emotions over time. LSTM layers process this time-series data by maintaining long-term contextual information. Finally, the model includes fully connected layers that interpret the learned features and classify emotions. The last layer uses a softmax activation function to produce probabilities for each emotion class (Srinivas et al., 2024).

3.7 Model Evaluation

3.7.1 Confusion Matrix

A confusion matrix is typically structured as a table that compares the predicted emotional states against the actual emotional states. Each row of the matrix represents instances in a predicted class, while each column represents instances in an actual class.

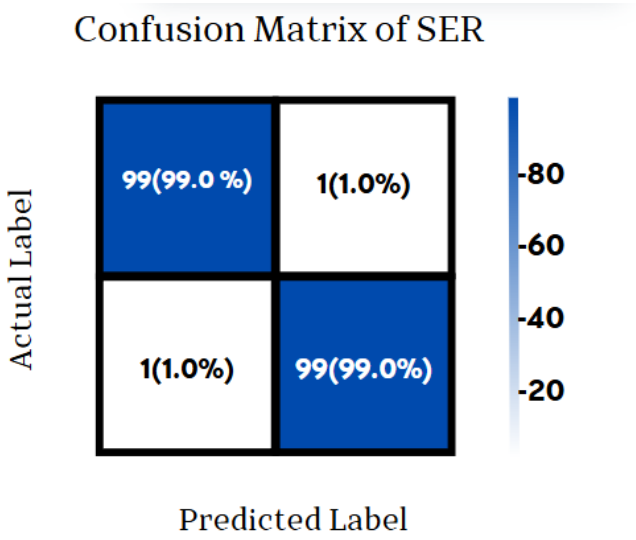


Figure 12: Example of Confusion Matrix of SER.

3.7.2 Evaluation Metric

Metric	Description	Formula
Precision	Precision refers to the ratio of correctly predicted positive observations to the total predicted positives	$Precision = \frac{TP}{TP + FP}$
Recall	Recall refers to the ratio of correctly predicted positive observations to all observations in the actual class	$Recall = \frac{TP}{TP + FN}$

Accuracy	Accuracy is a measure of the overall correctness of the model. It is the ratio of correctly predicted instances to the total instances	$Accuracy = \frac{TP + TP}{TP + FP + TN + FN}$
F1 Score	F1 score is the harmonic mean of precision and recall. It provides a balance between precision and recall	$F1\ Score = 2 * \frac{Precision * Recall}{Precisison + Recall}$

Table 4: Evaluation Metric of SER

From the table above, TP refers to TruePositive, TN refers to TrueNegative, FP refers to TruePositive and FN refers to FalseNegative.

3.8 Requirement Analysis

Requirement Analysis is a key process of gathering, understanding, and documenting the needs and constraints of a project or research. This analysis is to ensures that the necessary resources and tools are available to meet all the project objectives. Thus, in the research, the requirement analysis is performed to identifies the technical information and specifications such as the programming language and tools, libraries and hardware requirements that are needed for the implementation of SER model. This analysis helps in ensuring for smooth planning and implementation stages by considering both software and hardware requirements.

i) Programming Language, API and Platforms used for the project:

Requirement	Description
Programming Language	Python (Version 3.8 or above)
Application Programming Interface (API)	TensorFlow and Keras API for building deep learning models.

Tools/Platforms	Jupyter Notebook/Visual Studio Code

Table 5: Programming Language, API and Platforms for the implementation of SER model.

ii) Libraries for building the SER Model:

Library	Description
Librosa	For audio processing and feature extraction.
NumPy	For numerical computations and array manipulations.
Matplotlib	For visualizing data and performance metrics.
Scikit-learn	For model evaluation metrics (confusion matrix, accuracy, etc.).
Pandas	For data manipulation and analysis.
Seaborn	For plotting the confusion matrix and other performance visuals.

Table 6: List of Libraries for the developing of SER model.

iii) Hardware specification and minimum requirement to execute the program:

Hardware Specification	Minimum Requirement
Processor (CPU)	AMD Ryzen 3 3200U @ 2.60 GHz (or equivalent)
Graphics Card (GPU)	NVIDIA GTX 1060 (or equivalent)
(Random Access Memory) RAM	8 GB (Minimum) or above *Lower RAM causing longer waiting time in program execution and model training.
Storage	At least 250 GB of free space
Operating System	Windows 10 / Ubuntu 20.04 (or later)

Table 7: Hardware specification and requirements.

3.9 Summary

In this chapter, the methodology for building and evaluating a Speech Emotion Recognition (SER) model is explained. The proposed SER model is a hybrid model which consists of CNN and LSTM architectures. This chapter also explains about the use of the CREMA-D dataset for the project, along with preprocessing techniques such as data splitting, label encoding, and data augmentation to optimize the model training. Moreover, feature extraction is performed using Mel spectrograms. Finally, the chapter outlines the evaluation process which include metrics like precision, recall, accuracy, F1 score, and a confusion matrix to assess the model's performance.

Chapter 4 Implementation and Testing

4.1 Introduction

This chapter presents the implementation of the proposed Speech Emotion Recognition (SER) model. It outlines the complete experimental setup, starting with data collection of the CREMA-D and the data transformation steps, including data splitting, label encoding, and data augmentation. Feature extraction is performed using Mel spectrograms to convert audio signals into a visual representation for model input. The chapter also details the baseline experiment used for initial benchmarking and describes the development of the proposed 2D CNN-LSTM model architecture. Finally, it discusses the hyperparameter tuning process and the use of callback functions to optimize training performance.

4.2 Experimental Dataset

The dataset used in this research is CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset), an emotional multimodal actor dataset which designed for research in speech emotion recognition. It contains 7,442 clips from 91 actors with diverse ethnic backgrounds, ages ranging from 20 to 74, and a balanced gender distribution (48 male and 43 female) (Cao, Cooper, et al., 2014). In this dataset, actors delivered a selection of 12 sentences, each expressed with one of six emotions (anger, disgust, fear, happiness, neutral, and sadness) at four different intensity levels (low, medium, high, and unspecified). The distribution or number of each emotion is shown in Figure 13 below. Each of the emotions is represented in ANG, DIS, FEA, HAP, SAD and NEU respectively. The number of samples in ANG, DIS, FEA, HAP and SAD are the same which is 1271, while the number of samples for NEU emotion class is 1087.

```
df['label'].value_counts()
```

Python

```
label
ANG    1271
DIS    1271
FEA    1271
HAP    1271
SAD    1271
NEU    1087
Name: count, dtype: int64
```

Figure 13: Number of Each Emotions in CREMA-D.

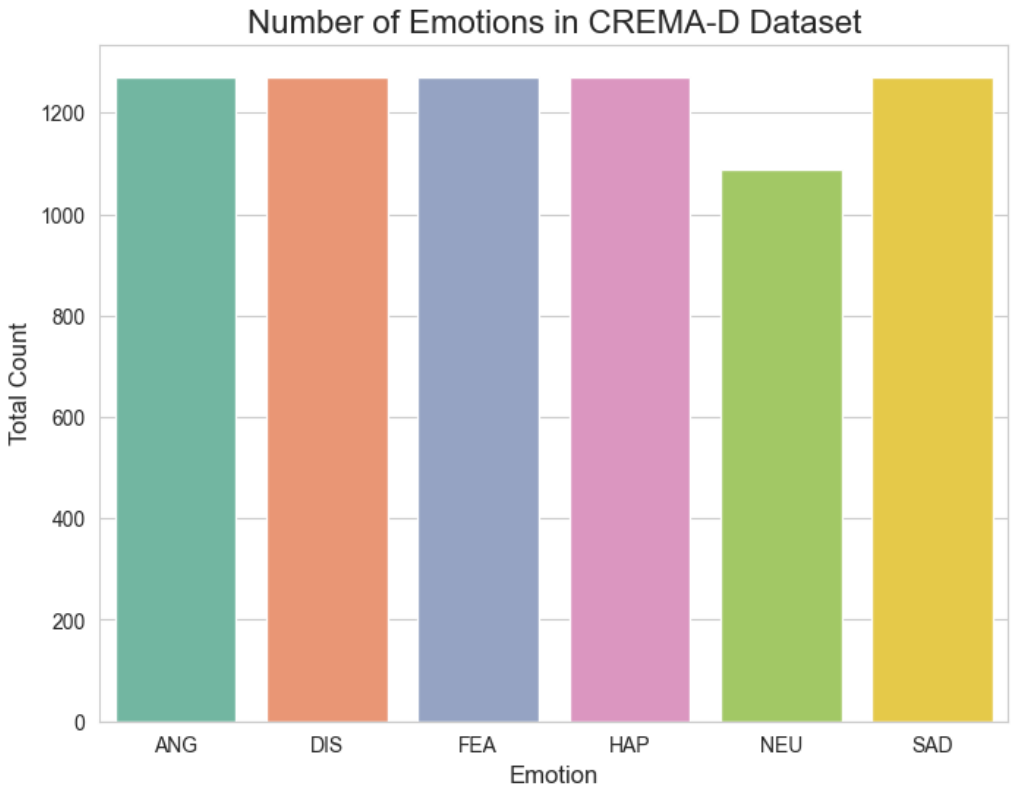


Figure 14: Bar Graph of Distribution of Every Emotion In CREMA-D.

To further visualize the sample distributions, Figure 14 above presents a bar graph showing the number of samples for each emotion. The graph clearly shows that all emotion categories are balanced, with the Neutral (NEU) emotion class has slightly fewer samples.

4.3 Data Transformation

4.3.1 Data Splitting

To prepare the dataset for the SER models training and evaluation, the CREMA-D dataset was randomly divided into three subsets which training set, validation set, and testing set.

```python
# Generate indices for splitting
X_indices = np.arange(len(X))

# First split: 70% Training, 30% Temporary (Validation + Testing)
indices_train, indices_temp, y_train, y_temp = train_test_split(
    X_indices, y, test_size=0.3, random_state=42, stratify=y
)

# Second split: 10% Validation, 20% Testing (from 30% temporary set)
indices_val, indices_test, y_val, y_test = train_test_split(
    indices_temp, y_temp, test_size=2/3, random_state=42, stratify=y_temp
)
```

Python

Figure 15: The Data Splitting Process for CREMA-D.

Based on the code snippet in Figure 15, the dataset was split using a two-step stratified sampling approach. First, 70% of the data was allocated for training, while the remaining 30% was set aside as a temporary subset for validation and testing. This was done using **'train_test_split()'** function with the stratify parameter to ensure that each emotion category remained proportionally represented. After that, the temporary 30% was further split into validation and testing sets in a 1:2 ratio which result in 10% of the total data for validation and 20% for testing. The number and proportion of each subset is shown in Figure 16 and Figure 17 below.

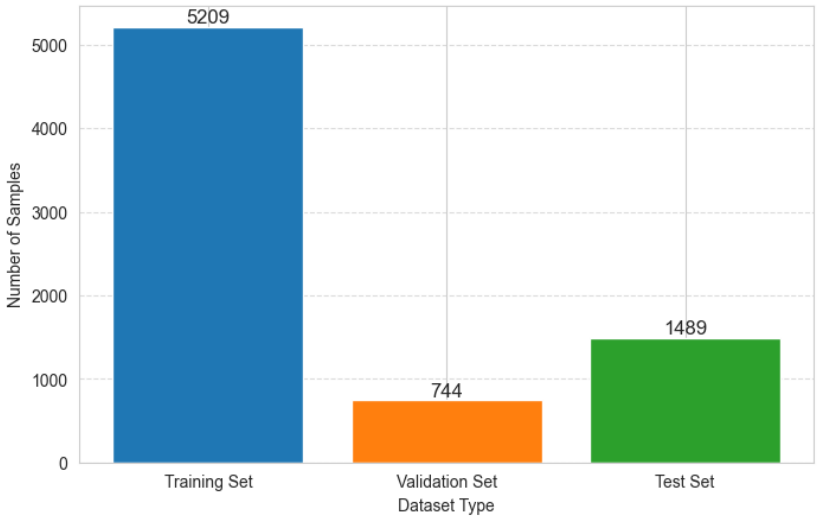


Figure 16: The number of each subset.

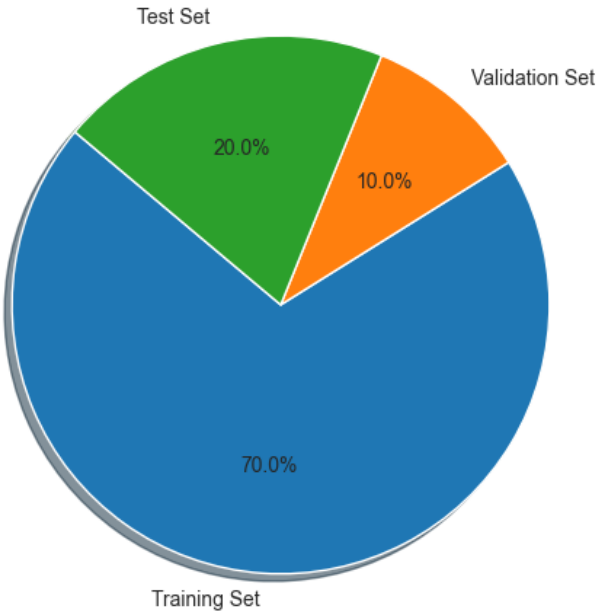


Figure 17: The Proportion of Each Subset.

4.3.2 Label Encoding

```
le = LabelEncoder()
y = utils.to_categorical(le.fit_transform(y))
y
```

Python

```
array([[1., 0., 0., 0., 0., 0.],
       [0., 1., 0., 0., 0., 0.],
       [0., 0., 1., 0., 0., 0.],
       ...,
       [0., 0., 0., 1., 0., 0.],
       [0., 0., 0., 0., 1., 0.],
       [0., 0., 0., 0., 0., 1.]])
```

Figure 18: Label Encoding Function

```
le_name_mapping = dict(zip(le.classes_, utils.to_categorical(le.transform(le.classes_))))
le_name_mapping
```

Python

```
{'ANG': array([1., 0., 0., 0., 0., 0.]),
 'DIS': array([0., 1., 0., 0., 0., 0.]),
 'FEA': array([0., 0., 1., 0., 0., 0.]),
 'HAP': array([0., 0., 0., 1., 0., 0.]),
 'NEU': array([0., 0., 0., 0., 1., 0.]),
 'SAD': array([0., 0., 0., 0., 0., 1.])}
```

Figure 19: A Dictionary which Mapping Each Class Label to its One-Hot Encoded Vector.

As shown in Figure 18 above, Label Encoding is performed using **'LabelEncoder()'**, which converts categorical class labels such as 'ANG', 'DIS', and 'FEA' into numerical values. After that, **'utils.to_categorical()'** is applied to the encoded labels to transform them into one-hot vectors where each class is represented as a binary array with a single 1 in the corresponding class position and 0s elsewhere. Additionally, a dictionary **'le_name_mapping'** is created to map each class label to its corresponding one-hot encoded vector.

4.4 Data Augmentation

When training and evaluating a machine learning model, a small training dataset can directly lead to deterioration in the trained model. Thus, increasing the number of training data through the data augmentation can help the model to learn more useful features and to avoid under-fitting during training (Pan & Wu, 2023).

Data augmentation is a technique to artificially increases the size of training dataset by creating modified versions of the original audio samples (Atmaja & Sasou, 2022). In this project, few data augmentation techniques are used in the model implementation including Additive White Gaussian Noise (AWGN), Time Shifting, and Pitch Shifting as shown in Figure 20 below.

```python
def awgn(data):
    noise_amp = 0.055 * np.random.uniform() * np.amax(data)
    data += noise_amp * np.random.normal(size=data.shape[0])
    return data
def time_shift(data, sr=44100, shift_limit=1):
    shift_amt = int(random.random() * shift_limit * len(data))
    return np.roll(data, shift_amt)
def pitch(data, sr=44100, pitch_steps=2):
    return librosa.effects.pitch_shift(data, sr=sr, n_steps=pitch_steps)
```

✓ 0.0s Python

Figure 20: Data Augmentation Techniques.

First data augmentation technique used is Additive White Gaussian Noise (AWGN). AWGN is a technique that adds random noise to the speech signal by generating a small amplitude noise component and adding it to the original waveform (Atmaja & Sasou, 2022). In this research, the noise ratio is set to 0.055 multiplied by a random uniform factor. The process of adding noise to the original speech signal involves two main steps. First, the maximum amplitude of the input signal is multiplied by the noise ratio and a random value between 0 and 1 to calculate the noise amplitude. Then, a Gaussian noise signal is generated using this noise amplitude and is added to the original speech waveform.

Next, the technique applied is **'time_shift'** which refers to Time Shifting function. Time Shifting involves the process of shifting the original audio signal along the time axis by a certain amount (Tao et al., 2022). In this research, a shift limit is defined relative to the length of the signal. A random shift amount is then calculated by multiplying this limit with a randomly generated value between 0 and 1. This shift value will determine how far the waveform will be shifted and altering the temporal position of the audio content without changing its structure.

Finally, the data augmentation technique used is **'pitch'** which refers to Pitch Shifting. Pitch Shifting is a function that modifies the pitch of the speech without affecting the duration of the speech (Amjad et al., 2022). This research utilizes the **'librosa.effects.pitch_shift()'** function to modify the pitch of the original speech signal without altering its duration. The process involves applying a Short-Time Fourier Transform (STFT) to convert the time-domain

signal into the frequency domain. Then, the frequency components of the signal are shifted by 2 semitones which is the number of steps to raise for the pitch. After that, the signal is reconstructed back into the time domain using an inverse STFT, resulting in a pitch-altered version of the original audio while maintaining its overall temporal structure. Figures below visualize each data augmentation technique used in the research.

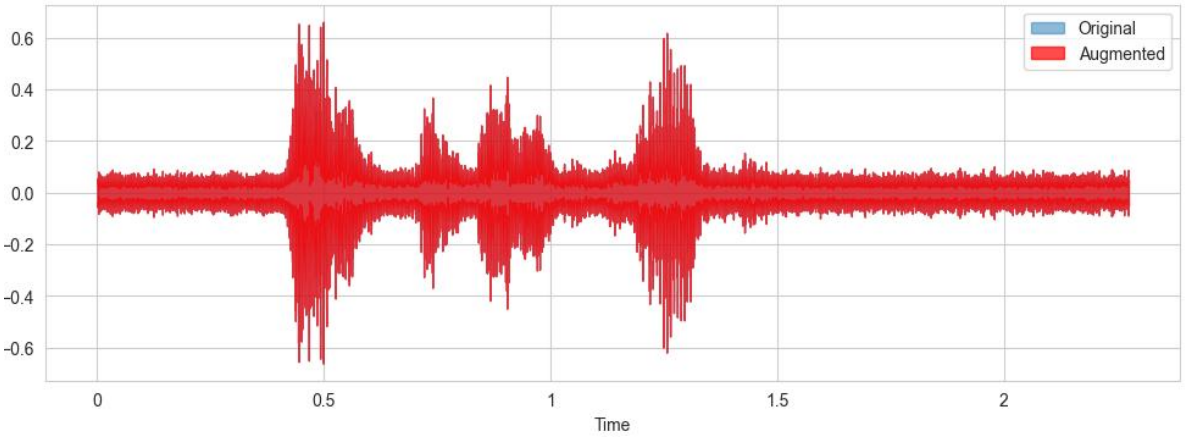


Figure 21: AWGN Augmentation Technique.

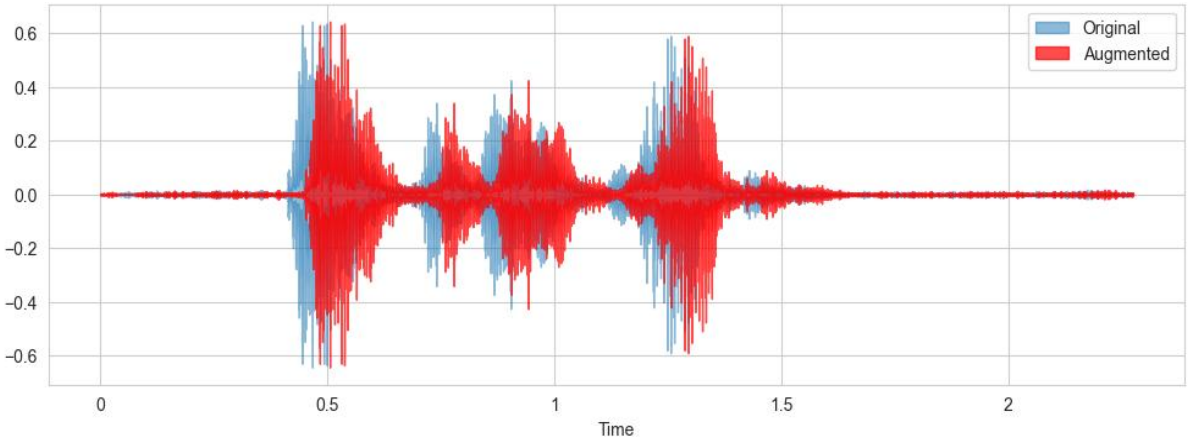


Figure 22: Time Shifting Augmentation Technique.

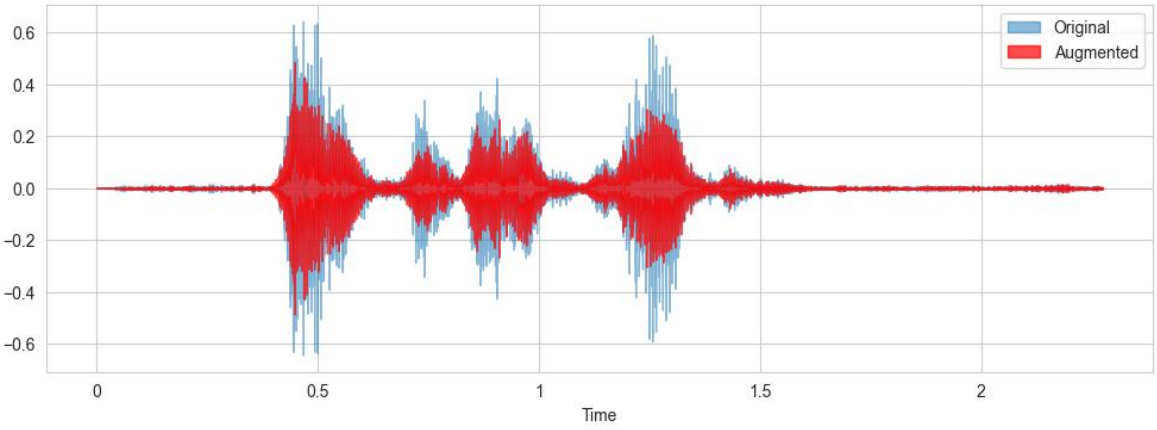


Figure 23: Pitch Shifting Augmentation Technique.

```python
def data_augment(X, y, aug_techniques=(awgn, pitch, time_shift)):
    awgn, pitch, time_shift = aug_techniques
    X_aug = []
    y_aug = []

    for i in range(len(X)):
        # Adding Original data
        X_aug.append(X[i])
        y_aug.append(y[i])

        # Adding AWGN and Time Shift
        X_aug.append(awgn(time_shift(X[i])))
        y_aug.append(y[i])

        # Applying Pitch and Time Shift
        X_aug.append(pitch(time_shift(X[i])))
        y_aug.append(y[i])


    X_aug = np.array(X_aug, dtype='object')
    y_aug = np.array(y_aug)

    return X_aug, y_aug
```

✓ 0.0s Python

Figure 24: Data Augmentation Function.

As shown in Figure 24 above, the '**data_augmentation()**' function accepts a set of original audio signals, *X* and their corresponding labels, *y*, along with three augmentation techniques which are Additive White Gaussian Noise (AWGN), pitch shifting, and time shifting.

This Data Augmentation Function involves three main steps for each audio sample. First, the original audio signal is retained to preserve the baseline data. Next, the audio signal undergoes time shifting followed by the addition of additive white Gaussian noise (AWGN). The augmented version of signal is then added to the dataset with its original label. Finally,

another variation is created by applying time shifting followed by pitch shifting, which is also appended to the dataset. This process generates two augmented samples for every original audio and effectively tripling the training set size from 5,209 to 15,627 samples as shown in Figure 25 below. The final augmented dataset is returned as a NumPy array to maintain the correct pairing between each transformed signal and its corresponding label.

```
X_aug_train, y_aug_train = data_augment(X_train, y_train)
✓ 12m 50.1s                                                    Python

print("Number of augmented training samples:", len(X_aug_train))
print("Number of corresponding labels:", len(y_aug_train))
✓ 0.0s                                                         Python

Number of augmented training samples: 15627
Number of corresponding labels: 15627
```

Figure 25: The Number of the Training Set After Data Augmentation..

4.5 Mel Spectrogram (Feature Extraction)

In speech emotion recognition (SER), feature extraction is a technique used to transforms raw audio signals into structured representations that are more suitable for machine learning models. Raw audio signals are usually high-dimensional and contain redundant or irrelevant information. Feature extraction addresses this by capturing the important vocal characteristics in the audio, such as pitch, tone, energy, and frequency patterns which are useful for recognizing different emotional states.

In this research, one feature extraction technique that has been employed is Mel Spectrogram. Mel Spectrogram is a time-frequency representation of audio that maps power spectrogram frequencies to the Mel scale, which is designed to mimics the way humans perceive sound. Mel spectrograms provide a visual representation of sound signals, where time

is displayed on the horizontal axis, frequency on the vertical axis, and amplitude intensity is represented by colour.

This technique begins with the Short-Time Fourier Transform (STFT) to break down audio into localized frequency components over time. Following this, Mel filter banks are applied. Mel filter banks consist of triangular-shaped filters that are spaced unevenly with more filters concentrated at lower frequencies. This design mimics how the human ear perceives sound and being more sensitive to lower frequencies. The resulting magnitudes are then converted into a logarithmic decibel (dB) scale to better align with human perception of loudness. Mathematically, the Mel scale is derived from linear frequencies using the formula shown in Equation 7 below.

$$f_{mel} = 2595\, log_{10}(1 + \frac{f}{700})$$

Equation 7: Mel Scale Formula (Abdelhamid et al., 2022).

4.5.1 Mel Spectrogram Function

```python
# Function to compute Mel Spectrogram
def mel_spectrogram(X, sr=44100, spec_shape=(128, 251)):
    X_mel = []
    for data in X:
        # Convert the raw waveform into a time-frequency representation
        sgram = librosa.stft(data)

        # Get the magnitude of frequencies
        sgram_mag, _ = librosa.magphase(sgram)

        # Convert the magnitude spectrum into a Mel spectrogram
        mel_scale_sgram = librosa.feature.melspectrogram(S=sgram_mag, sr=sr)

        # Convert the Mel spectrogram to the logarithmic decibel (dB) scale
        mel_sgram = librosa.amplitude_to_db(mel_scale_sgram, ref=np.min)

        # Resize the spectrogram to the specified shape
        img = Image.fromarray(mel_sgram)
        img = np.array(img.resize(spec_shape, resample=Image.LANCZOS))

        X_mel.append(img)

    return np.array(X_mel)
```

Python

Figure 26: Mel Spectrogram Function Used in the Research.

Based on the Mel Spectrogram function **'mel_spectrogram()'** shown in Figure 26 above, for each input audio sample, the function first applies the Short-Time Fourier Transform (STFT) using ***'librosa.stft()'*** to transform the waveform into a time-frequency domain. Then, it extracts the magnitude spectrum through **'librosa.magphase()'**. This magnitude spectrum is mapped onto the Mel scale using **'librosa.feature.melspectrogram()**' to simulate human auditory perception by emphasizing perceptually relevant frequency components. Next, the Mel spectrogram is converted to a logarithmic decibel (dB) scale through **'librosa.amplitude_to_db()'** to enhance the visibility of subtle features in the spectrogram. Finally, each Mel spectrogram is resized to a fixed dimension using the LANCZOS resampling method to ensure consistency in input shape for the SER models.

The visual representations of few emotion samples from both the original and augmented data, after applying Mel spectrogram feature extraction are shown in the figures below.

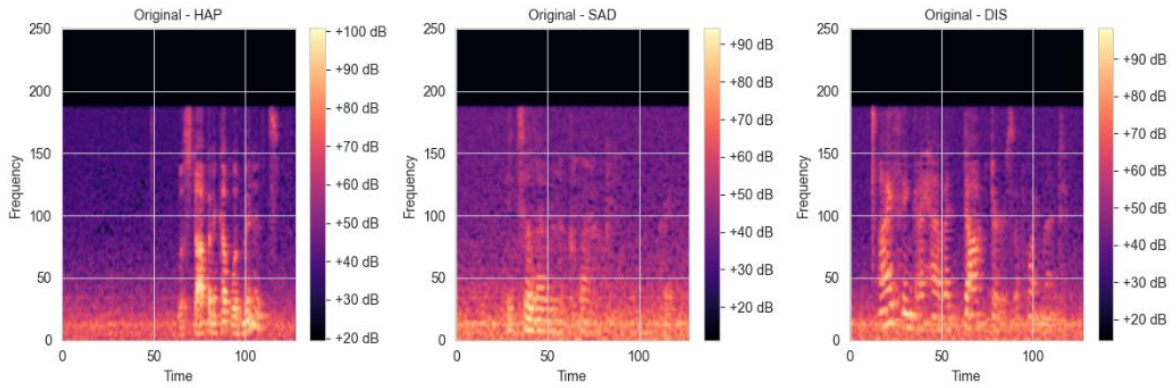


Figure 27: The Images of Original Dataset (Without Augmentation)

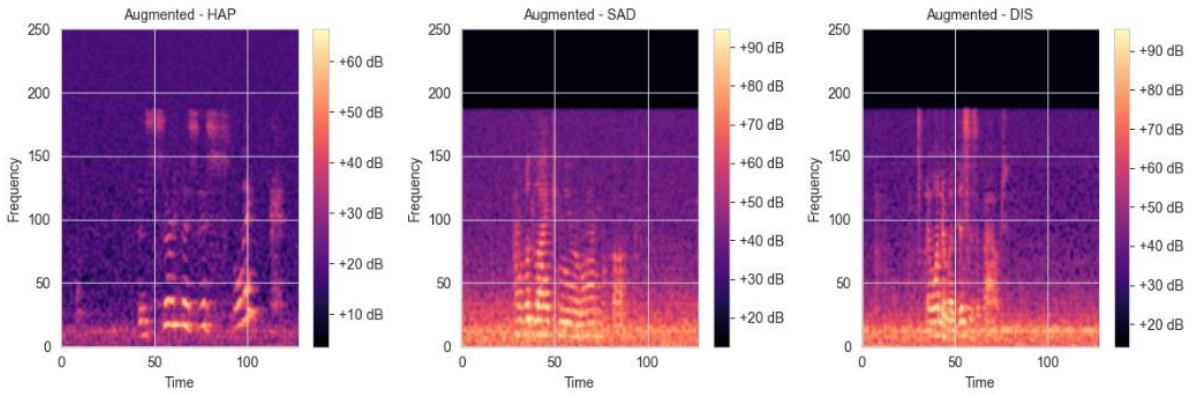


Figure 28: The Images of Augmented Dataset.

4.6 Baseline Experiment

In this study, a baseline Speech Emotion Recognition (SER) model is implemented using a single deep learning architecture which is Two-Dimensional Convolutional Neural Networks (2D CNNs). This baseline model is trained and evaluated on the same dataset as the proposed hybrid 2D CNN–LSTM model. This model will serve as a comparative reference point to assess the effectiveness of integrating temporal modelling components such as Long Short-Term Memory (LSTM) architecture. A baseline experiment refers to a foundational study that serves as a reference point for comparison (Wall & Horák, 2007). It establishes a standard against which a more advanced SER model can be evaluated to determine improvements or changes. In this case, it helps to determine whether integrating another deep learning architectures, such as LSTM layers will enhances the model's performance in recognizing emotions from speech or not. The illustration and summary of the architecture of the CNN model is shown in figure and table below.

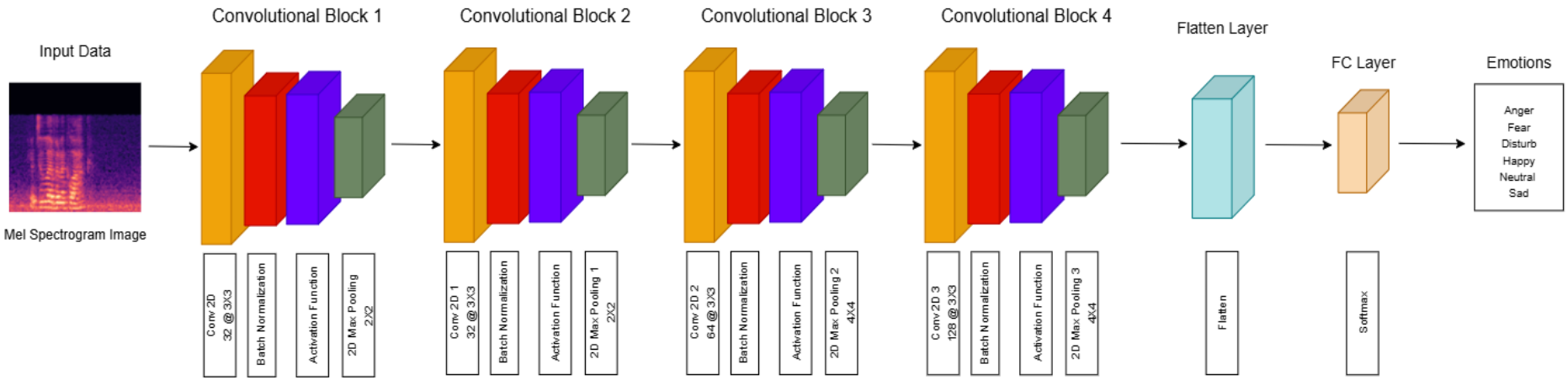


Figure 29: Illustration of 2D CNN Model (Baseline Experiment)

Layer Name	Layer types	Output Shape	Parameters
Convolutional Block 1	conv2d (Conv2D)	(None, 128, 251, 32)	320
	batch_normalization (BatchNormalization)	(None, 128, 251, 32)	128
	activation (Activation)	(None, 128, 251, 32)	0
	max_pooling2d (MaxPooling2D)	(None, 64, 125, 32)	0
Convolutional Block 2	conv2d_1 (Conv2D)	(None, 64, 125, 32)	9248
	batch_normalization_1 (BatchNormalization)	(None, 64, 125, 32)	128
	activation_1 (Activation)	(None, 64, 125, 32)	0
	max_pooling2d_1 (MaxPooling2D)	(None, 32, 62, 32)	0
Convolutional Block 3	conv2d_2 (Conv2D)	(None, 32, 62, 32)	18,496
	batch_normalization_2 (BatchNormalization)	(None, 32, 62, 32)	256
	activation_2 (Activation)	(None, 32, 62, 32)	0
	max_pooling2d_2 (MaxPooling2D)	(None, 8, 15, 64)	0
Convolutional Block 4	conv2d_3 (Conv2D)	(None, 8, 15, 64)	73,856
	batch_normalization_3 (BatchNormalization)	(None, 8, 15, 64)	512
	activation_3 (Activation)	(None, 8, 15, 64)	0
	max_pooling2d_3 (MaxPooling2D)	(None, 2, 3, 128)	0
Flatten Layer	flatten (Flatten)	(None, 768)	0
FC Layer	dense (Softmax)	(None, 6)	6

Table 8: The Architecture of CNN Model (Baseline Experiment).

The baseline architecture consists of four convolutional blocks (often known as Local Feature Learning Blocks or LFLBs), with of each of its consisting of a 2D convolutional layer with a kernel size of 3×3, followed by batch normalization, a Rectified Linear Unit (ReLU) activation function, and a 2D max-pooling layer. The first and second convolutional layers employ 32 filters and apply 2×2 max-pooling, while the third and fourth layers increase the filter size to 64 and 128, respectively. The latter two layers use a 2×4 max-pooling configuration. The architecture is designed to extract hierarchical time-frequency features from Mel spectrograms which will serve as the input representations of the speech signals.

Each convolutional layer in the model applies a set of learnable filters (or kernels) that slide over the input spectrogram image to extract meaningful local patterns. These filters

perform convolution operations including computing the dot product between the filter weights and localized patches of the input data. This will be resulting in the generation of feature maps. The 2D convolution operation at a given position (i, j) in the input spectrogram x, using a kernel w, can be mathematically represented as:

$$z(i,j) = \sum_{\{s=-a\}}^{a} \sum_{\{t=-b\}}^{b} x(i+s, j+t) \cdot w(s,t)$$

Equation 8: The Fomula for Two-Dimensional Convolution Operation (Chauhan et al., 2021).

Here, a and b denote the half-size of the kernel in each spatial dimension. This operation produces a feature map that emphasizes salient local patterns across time and frequency dimensions.

Following each convolutional operation, a Batch Normalization (BN) layer is employed to stabilize and accelerate the training process by normalizing the intermediate feature maps. BN transforms each input x using the formula:

$$BN(x) = \gamma \cdot \left(\frac{x - \mu}{\sqrt{\{\sigma^2 + \epsilon\}}}\right) + \beta$$

Equation 9: The Formula to Compute Batch Normalization (Ioffe & Szegedy, 2015).

Where μ and σ^2 are the mean and variance of the batch, γ and β are learnable scale and shift parameters, and ϵ is a small constant to prevent division by zero (Ioffe & Szegedy, 2015). By keeping the mean activation close to zero and the variance close to one, batch normalization improves the model's stability, reduces internal covariate shift and enables faster convergence during training.

The output is then passed through an activation function, ReLU (Rectified Linear Unit) which introduces non-linearity into the network and can be defined as:

$$ReLU(x) = max(0, x)$$

Equation 10: ReLU Function (Bai, 2022).

This function allows the network to learn complex and high-dimensional feature representations that are important in emotion recognition.

After the activation stage, a 2D max-pooling layer is applied to perform non-linear down-sampling. This down-sampling will reduce the spatial resolution of the feature maps while preserving the most important features. This process improves computational efficiency and reduces the risk of overfitting by making the model less sensitive to small variations in the input. The max-pooling output in the l -th layer can be expressed as:

$$z_k^l = \max_{\forall p \in \Omega_k} z\frac{l}{p}$$

Equation 11: Max Pooling Operation (J. Zhao et al., 2019).

Where Ω_k represents the pooling region with index k, z_k^l and z_p^l represents the output and input features respectively (J. Zhao et al., 2019). This down-sampling step retains the most prominent patterns in the input while reducing dimensionality, making it easier for the subsequent layers to process high-level features.

The final convolutional block's output is flattened into a one-dimensional vector and passed to a fully connected dense layer to aggregates the learned feature representations. Finally, a softmax classifier is used at the top layer to output a probability distribution over the target emotion classes. The softmax function transforms the logits z_i into a normalized probability vector p_i:

$$p_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_i}}$$

Equation 12: Softmax Function (J. Zhao et al., 2019).

where n is the number of emotion categories. The class with the highest probability is selected as the predicted emotional state of the input speech segment.

4.7 Model Development and Implementation

4.7.1 2D CNN LSTM Architecture (The Proposed Model)

The proposed 2D CNN-LSTM model builds on the baseline CNN by adding a recurrent layer, specifically, LSTM architecture. While the initial part of the model which made up of four convolutional blocks has remained unchanged as in the baseline CNN. The changes come after these convolutional blocks where the Flatten and Dense layers replace have been replaced with a Reshape layer and LSTM layers. The illustration and summary of the proposed SER Model (CNN-LSTM) is shown in figure and table below.

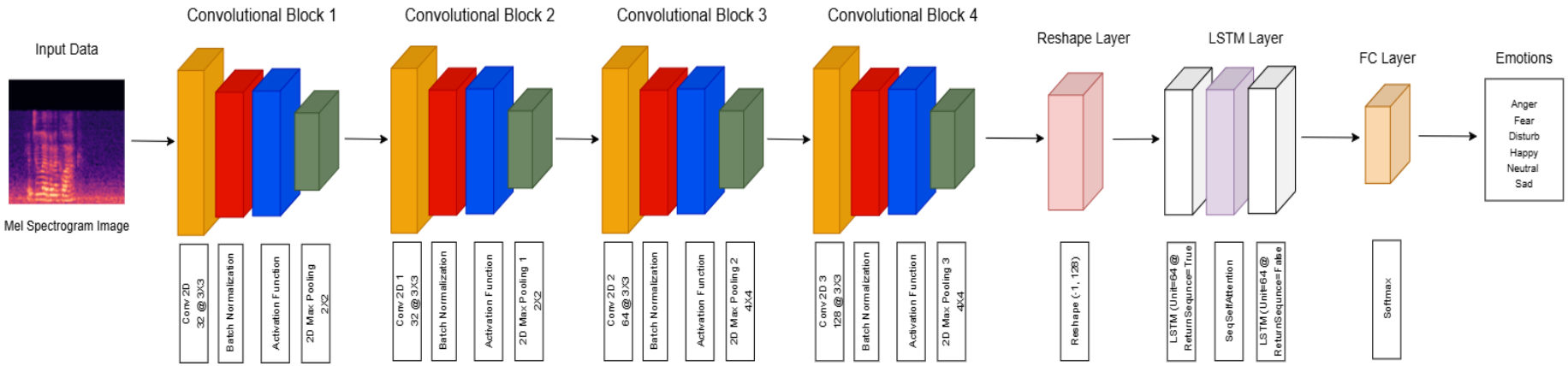


Figure 30: Illustration of Proposed 2D CNN LSTM Model.

Layer Name	Layer types	Output Shape	Parameters
Convolutional Block 1	conv2d (Conv2D)	(None, 128, 251, 32)	320
	batch_normalization (BatchNormalization)	(None, 128, 251, 32)	128
	activation (Activation)	(None, 128, 251, 32)	0
	max_pooling2d (MaxPooling2D)	(None, 64, 125, 32)	0
Convolutional Block 2	conv2d_1 (Conv2D)	(None, 64, 125, 32)	9248
	batch_normalization_1 (BatchNormalization)	(None, 64, 125, 32)	128
	activation_1 (Activation)	(None, 64, 125, 32)	0
	max_pooling2d_1 (MaxPooling2D)	(None, 32, 62, 32)	0
Convolutional Block 3	conv2d_2 (Conv2D)	(None, 32, 62, 32)	18,496
	batch_normalization_2 (BatchNormalization)	(None, 32, 62, 32)	256
	activation_2 (Activation)	(None, 32, 62, 32)	0
	max_pooling2d_2 (MaxPooling2D)	(None, 8, 15, 64)	0
Convolutional Block 4	conv2d_3 (Conv2D)	(None, 8, 15, 64)	73,856
	batch_normalization_3 (BatchNormalization)	(None, 8, 15, 64)	512
	activation_3 (Activation)	(None, 8, 15, 64)	0
	max_pooling2d_3 (MaxPooling2D)	(None, 2, 3, 128)	0
Reshape Layer	reshape (Reshape)	(None, 6, 128)	0
LSTM Layer	lstm (LSTM)	(None, 6, 64)	49,408
	seq_self_attention (SeqSelfAttention)	(None, 6, 64)	4161
	Lstm_1 (LSTM)	(None, 64)	33024
FC Layer	dense (Softmax)	(None, 6)	6

Table 9: The Architecture of CNN-LSTM Model (Proposed Model).

After the fourth convolutional block, instead of flattening the output into a one-dimensional (1D) feature vector, the output tensor is reshaped into a three-dimensional (3D) format of shape (batch size, time steps, features). This restructuring is necessary to prepare the feature maps for sequential modelling using Long Short-Term Memory (LSTM) architecture. The reshaping process preserves the temporal structure inherent in the spectrogram representation where the vertical axis corresponds to frequency and the horizontal axis

represents time. By treating one of the spatial dimensions as the temporal axis, the model can interpret the data as a time-ordered sequence of feature vectors.

The reshaped feature sequence is then fed into a stacked LSTM architecture for temporal sequence modelling. The first LSTM layer is configured with **'return_sequences=True'** to ensure that it outputs a complete sequence of hidden states corresponding to each time step. This configuration enables the model to retain contextual information and capture temporal dependencies across the duration of the speech input.

After that, to further enhance the model's ability to focus on emotionally salient segments, a self-attention mechanism is applied to the outputs of the first LSTM layer. This mechanism assigns different importance to each time step by computing a relevance score e_t for each hidden state h_t using a trainable scoring function f, typically implemented as a single-layer feedforward neural network:

$$e_t = f(h_t)$$

Equation 13: The Formula to Calculate Attention Score at Time Step t (Bahdanau et al., 2014).

These scores are then normalized via a softmax function to produce attention weights, α_t:

$$\alpha_t = \frac{\exp(e_t)}{\sum_{\{j=1\}}^{T'} \exp(e_j)}$$

Equation 14: The Formula to Calculate the Normalized Attention Weight α_t (Bahdanau et al., 2014).

Using these weights, a context vector c is computed as a weighted sum of the LSTM hidden states:

$$c = \sum_{t=1}^{T'} \alpha_t h_t$$

Equation 15: The Formula to Compute the Context Vector. C (Bahdanau et al., 2014).

The resulting context vector effectively encapsulates the most emotionally relevant temporal features from the input speech sequence. This allows the model to selectively focus on the most informative time frames.

Following the attention layer, the sequence is passed to a second LSTM layer configured with **'return_sequences=False'** to compress the time-distributed information into a single context vector. This vector summarizes the global temporal dynamics of the entire utterance, capturing both short- and long-range dependencies. The final context vector is then passed through a fully connected dense layer, followed by a softmax classifier that outputs a probability distribution over the predefined emotion classes.

By replacing the flattening mechanism with temporal modelling through LSTM layers and integrating an attention mechanism, the proposed CNN-LSTM model has effectively combined spatial feature extraction via convolutional layers with sequential feature learning as it captures both the local spectral patterns, and the temporal evolution of emotional cues present in speech signals.

4.8 Hyperparameter Tuning and Callback Functions

```
# Train Model
batch_size = 16
epochs = 50
```

Figure 31: The Hyperparameter Tuning for Model Training.

In this study, few hyperparameters were used to optimize the training process of both SER models. A batch size of 16 is being chosen and used for model training. This choice reflects a balance between computational efficiency and generalization performance while compared to higher batch size value. A smaller batch sizes often lead to more frequent weight updates and improved model adaptability. The model was also trained for 50 epochs, with each epoch representing a full pass through the training dataset. This duration is considered sufficient for the models to learn relevant patterns while also enable regular monitoring of validation performance.

```
early_stopping = EarlyStopping(monitor='val_loss', patience=7, restore_best_weights=True, mode = 'min')
checkpoint_rpcnn = ModelCheckpoint('model_rpcnn.keras', monitor='val_loss', save_best_only=True, verbose=1)
checkpoint_rpcnnlstm = ModelCheckpoint('model_rpcnnlstm.keras', monitor='val_loss', save_best_only=True, verbose=1)
lrs = ReduceLROnPlateau(monitor='val_loss', factor=0.2, patience=5, verbose=1, min_lr= 0.000001, mode = 'min')
```

Python

Figure 32: The Callback Functions Used for Models Training.

Several callback functions were also being implemented to improve the models' training process. Early stopping was employed to monitor the validation loss, with training process would be terminated if no improvement of validation loss was observed over seven consecutive epochs. This callback also restored the model weights from the epoch with the best validation performance. Model checkpointing was used to save the model with the lowest validation loss, ensuring that only the best-performing version of the model was retained

‘save_best_only=True’. Additionally, a learning rate reduction was applied through **‘ReduceLROnPlateau’**, which reduced the learning rate by a factor of 0.2 if the validation loss did not improve after five epochs, with a minimum threshold set to avoid excessively small learning rates. All callbacks were set with verbose=1 to provide real-time feedback during training.

4.9 Summary

This chapter presented a comprehensive overview of the implementation processes for the Speech Emotion Recognition (SER) models. It detailed the data transformation processes, including label encoding, data augmentation, train-test data splitting, and the conversion of audio signals into Mel spectrogram representations for model input. Two SER models were implemented: a 2D CNN model as the baseline and a 2D CNN-LSTM model as the proposed architecture. Both models were trained and evaluated using the CREMA-D dataset. The chapter also described the architectural design of each model and outlined the training procedures employed to optimize their performance.

Chapter 5 Result and Analysis

5.1 Introduction

This chapter presents and analyses the experimental results obtained from both the baseline and proposed speech emotion recognition (SER) models. The models' performance is assessed using key evaluation metrics, including accuracy, precision, recall, F1-score, and confusion matrix. This chapter also details the accuracy trends on the validation and testing sets for both models. The classification performance is further examined through statistical metrics and confusion matrices. In addition, a benchmark study that compares the proposed model's performance with findings from existing studies to evaluate its effectiveness and highlight potential improvements in emotion recognition is also highlighted in this chapter.

5.2 SER Model Accuracy on Validation & Testing Sets

SER Model Type	SER Model Architecture	Validation Accuracy (%)	Testing Accuracy (%)
Baseline Model	2D CNN	62	58
Proposed Model	2D CNN LSTM	60	59

Table 10: The Accuracy of Baseline Model on Validation & Testing Sets.

The baseline 2D CNN model achieved a validation accuracy of 62% and a testing accuracy of 60%, whereas the proposed 2D CNN-LSTM model achieved a slightly lower validation accuracy of 60% but higher testing accuracy of 59%. While both models have achieved almost similar result in term of accuracy, the proposed 2D CNN-LSTM model demonstrates a slightly better accuracy in test set which indicating an improvement in model performance and generalization to the testing dataset.

5.2.1 Graphs of Accuracy and Loss for Baseline and Proposed Model During Training

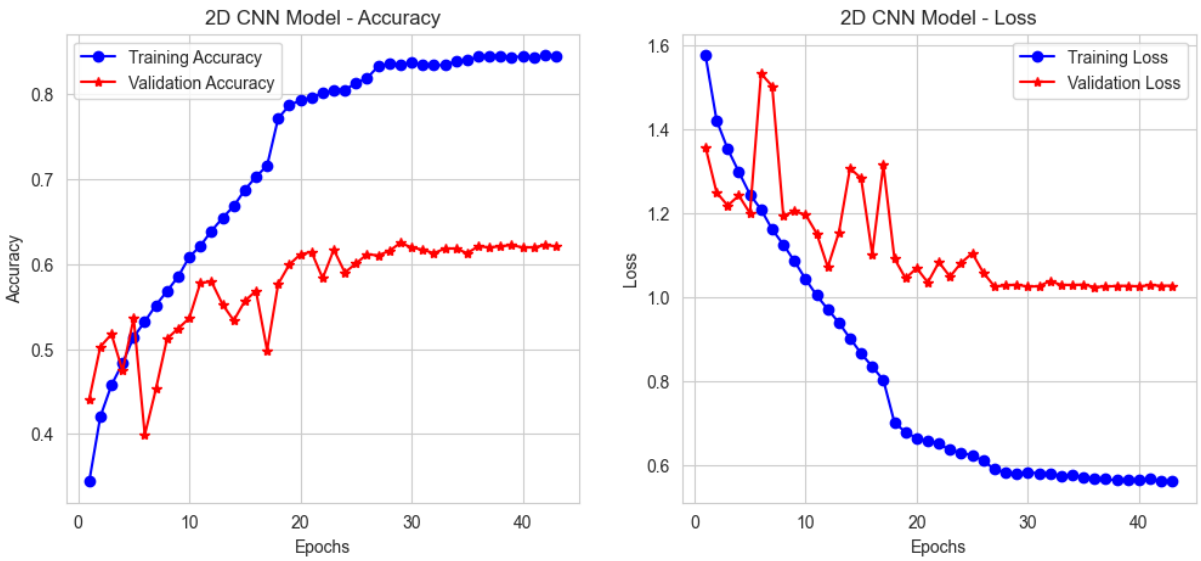


Figure 33: Training and Validation Graphs which showing for Accuracy and Loss for Baseline Model (2D CNN).

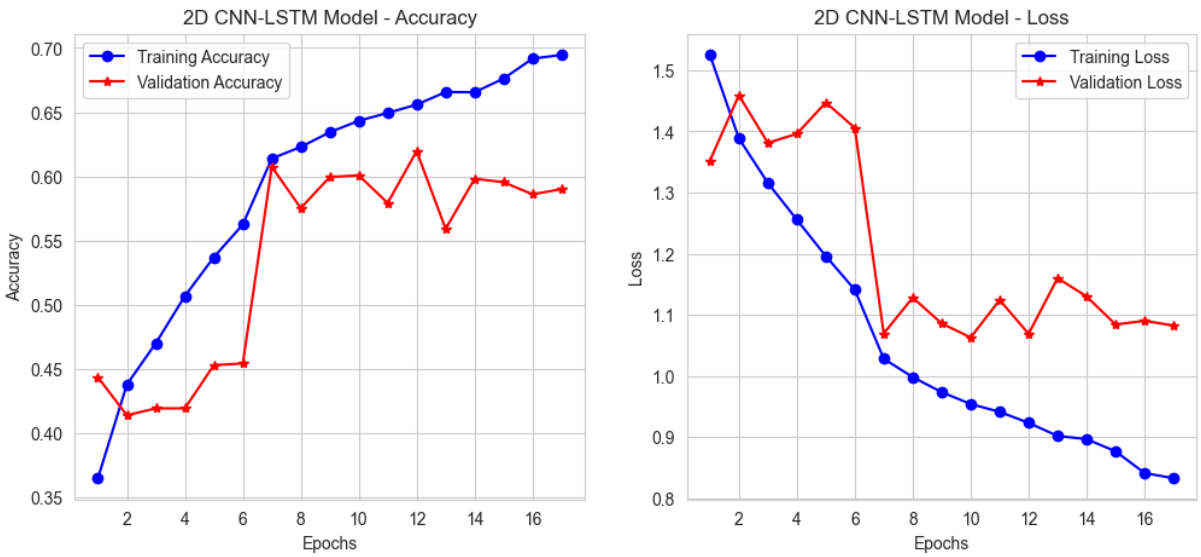


Figure 34: Training and Validation Graphs which showing for Accuracy and Loss for Proposed Model (2D CNN LSTM).

Based on Figure 33, the training accuracy of the baseline 2D CNN model steadily increases and saturates around 0.84, whereas the validation accuracy plateaus at approximately 0.62. This gap

suggests that the model is overfitting as it learns the training data well but does not perform as effectively on new and unseen data on the testing set. Meanwhile, the training loss keeps decreasing, while the validation loss settles near 1.1 after some early changes which also supports the overfitting observation.

For the training process of 2D CNN-LSTM model shown in Figure 34, training accuracy improves progressively and reaching approximately 0.69 at epochs 17. However, validation accuracy becomes unstable after the seventh epoch and stays between 0.57 and 0.62 without significant improvement. This inconsistency, along with the gap between training and validation accuracy also indicates this model may seem to be overfitting. While the training loss continues to drop, the validation loss remains uneven and does not decrease much which shows that the model may struggle to generalize well.

5.2.2 Precision, Recall, F1-Score and Confusion Matrix

i) Experiment Result of Baseline Model (2D CNN)

	Precision	Recall	F1-Score	Support
ANG (0)	0.69	0.72	0.70	254
DIS (1)	0.55	0.56	0.56	255
FEA (2)	0.52	0.51	0.51	254
HAP (3)	0.58	0.55	0.56	254
NEU (4)	0.63	0.61	0.62	217
SAD (5)	0.54	0.56	0.55	255

Table 11: The Result of Precision, Recall & F1 Score for Baseline Model on Test Set.

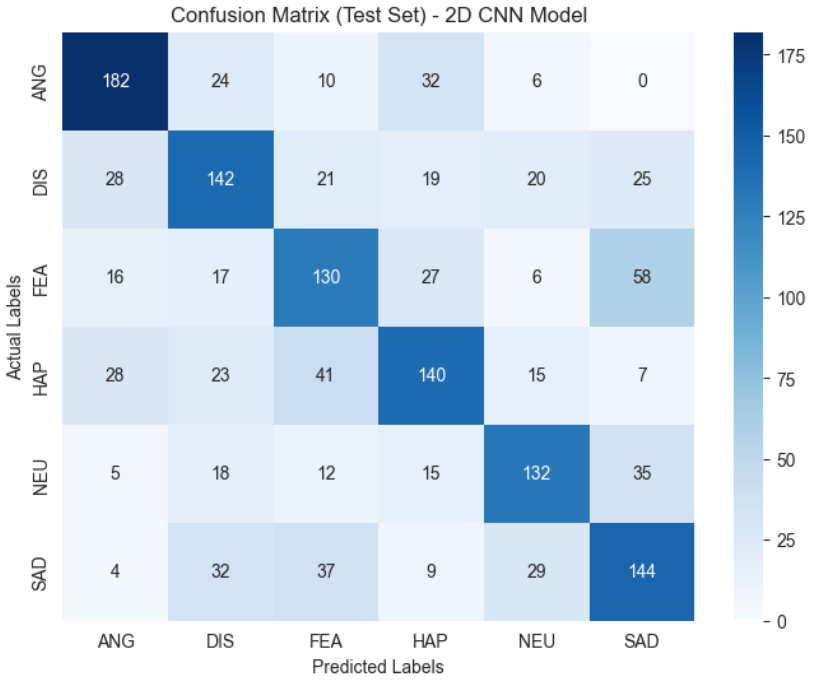


Figure 35: The Confusion Matrix for Baseline Model (2D CNN) on Test Set.

Based on Table 12 and Figure 36 above, baseline 2D CNN model shows average performance in recognizing all the emotion classes. ANG (Anger) is the best-recognized emotion, with an F1-score of 0.70. Out of 263 actual angry samples, 182 were correctly predicted. NEU (Neutral) is the second-best emotion recognized with an F1-score of 0.62, and 132 out of 208 samples were correctly classified.

Meanwhile, FEA (Fear) is the weakest-performing emotion, with the lowest F1-score of 0.51. The confusion matrix shows that only 130 out of 251 fear samples were correctly identified, while many were incorrectly predicted as SAD (Sadness), HAP (Happiness) and DIS (Disgust). These patterns highlight the model's difficulty in differentiating emotions with similar acoustic characteristics and features. The emotions Disgust, Happiness, and Sadness show moderate results, with F1-scores ranging between 0.55 and 0.56.

i) Experiment Result of Proposed Model (2D CNN LSTM)

	Precision	Recall	F1-Score	Support
ANG (0)	0.62	0.80	0.70	254
DIS (1)	0.56	0.35	0.43	255
FEA (2)	0.55	0.55	0.55	254
HAP (3)	0.56	0.57	0.56	254
NEU (4)	0.62	0.69	0.65	217
SAD (5)	0.62	0.61	0.62	255

Table 12: The Result of Precision, Recall & F1 Score for Proposed Model on Test Set.

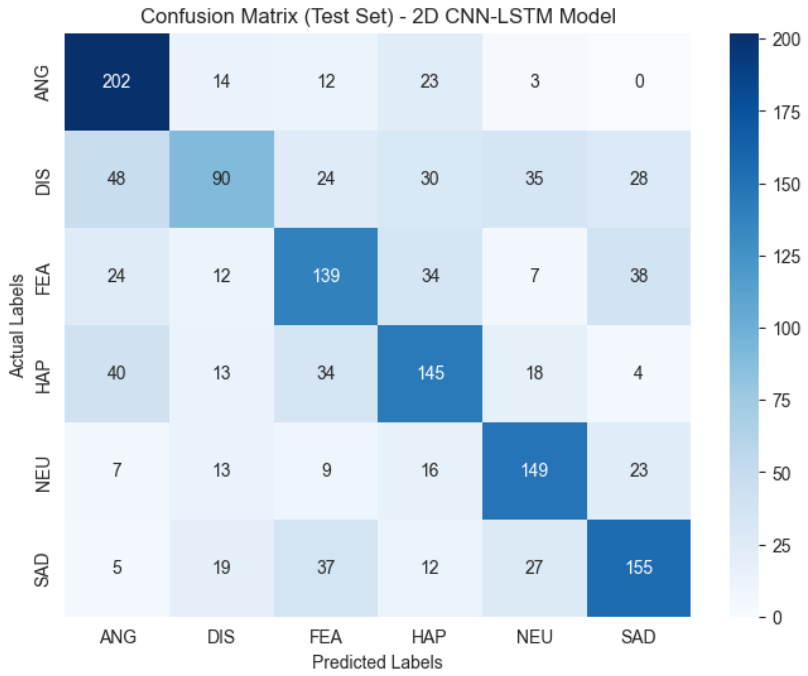


Figure 36: The Confusion Matrix for Proposed Model (2D CNN LSTM) on Test Set.

On the other hand, from the experiment result and confusion matrix shown above, proposed 2D CNN-LSTM model shows improved recognition for certain emotions compared to the baseline 2D CNN Model especially in recognizing Sadness emotion. Anger, Neutral, and Sadness are

the best-performing classes where these emotions achieved F1-scores of 0.70, 0.65, and 0.62, respectively. Anger has a high recall of 0.80, with 202 out of 326 samples correctly identified. However, its lower precision (0.62) suggests that many other emotions such as Disgust, Fear, and Happiness were incorrectly predicted as Anger.

Neutral emotion also improved, with 149 correct predictions out of 239 samples, and was mainly misclassified as Sadness. Sadness itself performed better than in the baseline model, showing a more balanced result with 155 correct predictions. On the other hand, Disgust emotion performed the worst, with an F1-score of 0.43. Only 90 out of 161 samples were correctly identified. This indicates the model struggled to learn different acoustic features for Disgust emotion.

Meanwhile, Fear and Happiness showed moderate performance, with F1-scores of 0.55 and 0.56, respectively. Fear was confused mostly with Sadness, Happiness, and Anger, while Happiness was often misidentified as Anger or Fear.

5.3 Experiment Summary

To conclude the findings from the experiment of both models, the proposed 2D CNN-LSTM model has performed slightly better than the baseline 2D CNN model with higher achieving accuracy and better generalization performance in recognizing new, unseen data. Although the changes may not be significant, the improvement in overall accuracy from 58% to 59% highlights the value of integrating temporal modelling such as LSTM architecture into speech emotion recognition (SER) systems.

While it can effectively capture localized spectral features, the baseline 2D CNN model showed limitations in recognizing emotions that require temporal context. For example, this model struggled to differentiate between Fear and Sadness emotions where this model has misclassifying around 58 fear samples as Sadness and 37 sadness samples as Fear. This confusion can be attributed to CNN's frame-independent processing as it fails to capture temporal acoustic features such as rhythm, pitch, and intonation patterns that change over the time (J. Zhao et al., 2019).

The proposed model addressed these limitations by integrating two LSTM layers along with a self-attention mechanism. This allowed the model to learn dependencies across time steps and focus most emotionally important segments of the audio. For example, the LSTM layers enabled the system to relate high-energy patterns with sustained intensity. This is shown as it led to reduction in misclassifying such as Happiness as Anger. On the other hand, the attention mechanism enhanced recognition of Neutral and Fear emotions by emphasizing consistent pitch or irregular voicing features, enhancing Neutral emotion's recall from 0.61 in baseline model to 0.69.

Moreover, stacked LSTMs also helped the model learn different time-based patterns with one layer focusing on short-term (syllabic) patterns and the other capturing longer

emotional changes (Yang et al., 2024). These enhancements contributed to improvement of F1-score across several emotion classes as shown in the Table below.

Emotion Class	F1-Score (Baseline Model)	F1-Score (Proposed Model)	Changes in F1-Score
SAD	0.55	0.62	+ 0.08
NEU	0.62	0.65	+ 0.03
FEA	0.51	0.55	+ 0.05

Table 13: Improvement in F1 Score.

Despite these improvements, performance for the Disgust emotion declined as F1 score dropped from 0.56 to 0.43 when compared to baseline experiment. This decrement is most likely due to the similar acoustic features with Anger emotions.

In summary, the proposed SER model achieved a 1% improvement in accuracy on new, unseen data, showing that it handles time-based patterns better than a standalone CNN model. Although the accuracy gain may seem little, the enhanced ability to distinguish emotions and generalize across varied inputs highlights the value of integrating LSTM architecture into CNN-based models.

5.4 Benchmark Study

A deeper analysis of proposed 2D CNN LSTM model has also been conducted by comparing it to other Speech Emotion Recognition Models. This analysis is important to measure the effectiveness and to identify few limitations of the proposed model. The table of comparison between previous studies that employing hybrid SER architecture is shown below.

Research Paper	Speech Emotion Recognition Model	Feature Extraction	Data Augmentation (Yes/No)	Data Splitting	Dataset(s)	Model Accuracy (%)
"CREMA-D: Improving Accuracy with BPSO-Based Feature Selection for Emotion Recognition Using Speech"- (Donuk, 2022).	CNN+BPSO+SVM (BPSO- Binary Particle Swarm Optimization)	MFCCs (Mel Frequency Cepstral Coefficients)	No	80% Training, 20% Testing.	CREMA-D	66.01
"Performance Improvement of Speech Emotion Recognition Systems by Combining 1D CNN and LSTM with Data Augmentation" - (Pan & Wu, 2023)	1D CNN LSTM	MFCC	Yes	81% Training, 9% Validation, 10% Testing.	1) RAVDESS 2) EMO-DB 3) IEMOCAP	95.52 (RAVDESS), 95.84 (EMO-DB), 96.21 (IEMOCAP)
"Speech emotion recognition using deep 1D & 2D CNN LSTM networks" - (J. Zhao et al., 2019)	2D CNN LSTM	Log-Mel spectrogram	No	80% Training, 20% Testing	1) Berlin-EmoDB 2) IEMOCAP	95.33 (EMO-DB), 89.16 (IEMOCAP).
"Robust Speech Emotion Recognition Using CNN+LSTM Based on Stochastic Fractal Search Optimization Algorithm" - (Abdelhamid et al., 2022)	2D CNN-LSTM	Log-Mel spectrogram	Yes	64% Training, 16% Validation, 20% Testing.	1) RAVDESS 2) IEMOCAP 3) EMO-DB 4) SAVEE	99.47 (RAVDESS), 98.13 (IEMOCAP), 99.76 (EMO-DB), 99.50 (SAVEE)
Proposed Model	2D CNN LSTM	Mel Spectrogram	Yes	70% Training, 10% Validation, 20% Testing.	CREMA-D	59.1

Table 14: Benchmark Analysis and Comparison.

This integration of both CNN and LSTM architectures in the proposed model has leads to improved recall for certain emotions such as anger, neutral, and sadness, compared to CNN-only approaches. But although it has demonstrated improvements and strength when compared to the baseline model, this model has also shown clear limitations such as lower achieved accuracy when compared to previous studies.

For example, Donuk (2022) achieved 66.01% accuracy on the CREMA-D dataset using a CNN with a feature selection (BPSO) and SVM classifier. Meanwhile, Pan and Wu (2023) reported much higher accuracies with over 95% across several datasets by combining 1D CNN–LSTM models with data augmentation techniques. Similarly, Zhao et al. (2019) and Abdelhamid et al. (2022) achieved over 89% accuracy by using similar hybrid CNN LSTM models and optimization algorithms like Stochastic Fractal Search. Meanwhile, the proposed model achieved a test accuracy of only 59.1% with 880 test samples correctly predicted out of 1489, and about 609 samples are misclassified.

Emotion Label	Number of emotions that predicted wrongly
Angry (ANG)	124
Disgust (DIS)	71
Fear (FEA)	116
Happiness (HAP)	115
Neutral (NEU)	90
Sadness (SAD)	93

Table 15: Number of each emotion that predicted wrongly.

Emotion Label	Distribution of Actual Label (Reference)	Distribution of Predicted Emotion (Hypothesis)
Angry (ANG)	254	326
Disgust (DIS)	255	161
Fear (FEA)	254	255
Happiness (HAP)	254	260
Neutral (NEU)	217	239
Sadness (SAD)	255	248

Table 16: Comparison of Actual vs Predicted Emotion Label Distributions.

Emotion Label	Misclassification Percentage (%)
Angry (ANG)	38.04
Disgust (DIS)	44.10
Fear (FEA)	45.49
Happiness (HAP)	44.23
Neutral (NEU)	37.66
Sadness (SAD)	37.50

Table 17: Misclassification Percentage for Every Emotion Class.

While the proposed 2D CNN–LSTM model demonstrated improvements compared to a CNN-only baseline, several challenges and limitations were observed during evaluation which may lead it to achieve a lower accuracy when compared to previous study. The model seems to struggle particularly with recognizing the Disgust emotion. Analysis of the confusion matrix revealed that most Disgust samples were misclassified as other emotions where 48 samples were predicted as Anger, 35 as Neutral, 30 as Happiness, 28 as Sadness, and 24 as Fear. Only 90 Disgust samples were correctly classified out of 161 in Distribution of Predicted Emotion (Hypothesis) in the test set. This challenge may be attributed to the difficulty in identifying the Disgust emotion due to its subtle acoustic characteristics (Rezapour Mashhadi & Osei-Bonsu, 2023).

Moreover, similar like Disgust emotion, the model also showed high misclassification rates for Fear (45.49%) and Happiness (44.23%) emotions. Although the overall accuracy improved, Fear and Happiness samples continued to be misclassified frequently which indicating that the model facing difficulty in capturing certain prosodic and temporal features associated with these emotions. This confusion may be due to the overlapping acoustic patterns, as Fear, Happiness, and Sadness can sound similar, while Happiness was often mistaken for high-arousal emotions like Angry.

Dataset Type	Total Number of Audio Samples	Speech Intensity Level			
		Unspecified	Low	Medium	High
Training Set	5209	4248	319	319	323
Testing Set	1489	1213	99	89	88

Table 18: Audio Sample Distribution Across Training and Testing Sets by Intensity Level.

Emotion Class	Total Number of Audio Samples	Speech Intensity Level			
		Unspecified	Low	Medium	High
Disgust (DIS)	889	695	67	64	63
Fear (FEA)	890	711	54	61	64
Happiness (HAP)	890	699	62	61	68

Table 19: Emotion Class Distribution in Training Set by Speech Intensity Level.

Emotion Class	Total Number of Audio Samples	Speech Intensity Level			
		Unspecified	Low	Medium	High
Disgust (DIS)	255	207	14	19	15
Fear (FEA)	254	193	28	17	16
Happiness (HAP)	254	198	22	18	16

Table 20: Emotion Class Distribution in Testing Set by Speech Intensity Level.

A significant factor contributing to the relatively low accuracy and high misclassification rates observed for emotions such as Disgust, Fear, and Happiness in the

CREMA-D dataset is the clear imbalance in the distribution of speech intensity labels. Specifically, approximately over 80% of the audio samples across both training and testing sets as shown in Figure 16 are labelled as “Unspecified” intensity, while the remaining samples labelled as low, medium, and high intensity, each constitute only about 6 to 7% of the sets. For audio samples recorded as “Unspecified” intensity level, actors likely expressed emotion naturally without fixed intensity guidance. This strong imbalance in intensity representation limits the model’s ability to effectively learn the prosodic variations-particularly in energy and pitch-that correspond to different levels of emotional expressiveness.

Emotional arousal is closely associated with vocal intensity. For example, high-arousal emotions, such as anger and fear are generally characterized by increased amplitude and higher fundamental frequency, whereas low-arousal emotions, such as sadness and disgust tend to be quieter and more monotone. However, because the model is predominantly trained on speech samples with unspecified intensity, it tends to disregard these important acoustic cues which related to these emotions and may treating intensity as an irrelevant feature. As a result, the model fails to accurately capture the spectral-temporal patterns that differentiate emotions across different loudness levels. Instead, it overfits to the dominant unspecified category where it learns only the acoustic characteristics most common in mixed-intensity samples.

The relatively high number of unspecified-intensity speech during both training and evaluation biases the model toward representations that are insensitive to intensity variations. This bias inflates error rates on minority intensity subsets and obscures underlying weaknesses in the overall accuracy metric and performance generalization of the proposed model.

5.5 Discussion

Based on both models' evaluation and deeper analysis, there are few key strengths and limitations can be highlighted from this research. One of the key strength or advantages is the effective use of data augmentation and feature extraction technique, specifically Mel Spectrogram. By applying augmentation techniques such adding noise (AWGN), time shifting and pitch shifting, the training data of CREMA-D was expanded significantly and tripling the number of training samples available. This helped to avoid underfitting during the model training and improved its ability to generalize beyond the training set. Additionally, the use of Mel spectrograms as input features was particularly effective. These spectrograms represent speech in a way that closely matches how humans perceive sound especially in capturing important features like pitch and tone.

Another strength lies in the architecture of the proposed SER model itself. By combining convolutional neural networks (CNNs) with long short-term memory (LSTM) layers allowed the model to not only extract spatial features from the spectrograms but also capture the temporal flow of speech which is important for understanding different emotions. The addition of a self-attention mechanism further improved the model by helping it focus on the most emotionally relevant parts of each utterance. This combination led to better performance compared to a baseline CNN model with higher accuracy and improved F1-scores across most emotion categories.

However, the research has also revealed some limitations and disadvantages. Based on the experiment graph of training and validation, both the baseline CNN and proposed CNN-LSTM models showed signs of overfitting where they performed much better on training data than on validation data. This suggests that the models may have memorized specific samples

of the data rather than learning important features. Moreover, certain emotions like Disgust, Fear, and Happiness were often confused with other emotions and have higher misclassification rates. Although the CNN-LSTM model improved results, the overall accuracy of around 59% is still moderate compared to some previous studies that achieved higher accuracy.

Finally, the dataset itself also presents some limitations. While CREMA-D is a well-known and widely used dataset, it has an imbalance in emotion intensity labels with most of the samples annotated as "Unspecified." This dataset imbalance made it harder for the model to learn fine-grained differences between emotions that depend on subtle vocal changes, such as fear and disgust. Additionally, because the dataset consists of acted emotions, the expressions may sometimes be exaggerated or less natural than real-life speech which could affect how well the model performs in practical applications. Addressing these issues by using more natural datasets and balancing the data better could help improve future speech emotion recognition systems.

5.6 Summary

In this chapter, the performance of the proposed model is evaluated using key metrics, including accuracy, precision, recall, F1-score, and the confusion matrix, to assess its improvements over the baseline model. Additionally, a benchmark study is conducted to compare the proposed model's performance with previous studies, providing insights into its effectiveness and identifying potential limitations.

Chapter 6 Conclusion and Future Work

6.1 Introduction

In this chapter, the achievement of project questions and objectives, the limitations of the study, concluding insights, and potential directions for future work are discussed. These reflections are grounded in the findings from the model development and experimentation phases, which involved training and evaluating the proposed model on the CREMA-D dataset. Primary focus is given to the impact of dataset characteristics such as class imbalance and expressive variability on model performance. Furthermore, the potential applications of Speech Emotion Recognition (SER) in mental health monitoring are explored, emphasizing the importance of improving model accuracy for practical and real-world deployment.

6.2 Research Achievements

6.2.1 Summarization for Research Questions

Research Questions	Description
1. What are the recent trends in deep learning techniques applied in speech emotion recognition?	- This research question is answered through the comprehensive literature review on various deep learning architecture such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM), Transformer Model and Hybrid Model. This review focused the function and key strength of each architecture in the context of speech emotion recognition.
2. Which deep learning technique is the most suitable to be used for developing acoustic model and how this technique can recognize different emotions from CREMA-D dataset?	- This research question is answered by comparing previous studies in speech emotion recognition. The comparison focuses on factors such as the acoustic models employed (specifically deep learning techniques), the datasets used for training and testing and the accuracy achieved by each model. From this analysis, hybrid models utilizing CNN-LSTM architectures have demonstrated high accuracy in emotion recognition, typically ranging between 80% to 99%. Thus, this hybrid model

	is proposed for development and evaluation using the CREMA-D dataset.
3. How does the performance of the trained model on the CREMA-D dataset compare with existing models?	- This research question is answered through the evaluation of the both baseline and proposed acoustic models using key performance metric including Recall, Precision, Accuracy and F1-Score after being trained and tested on CREMA-D. After that, the proposed model which has achieving higher accuracy and better performance that baseline model is compared to SER models in previous studies to identify the model's limitations

Table 21: The Summary of Research Questions.

6.2.2 Summarization for Research Objectives

This project aimed to accomplish three specific objectives as outlined during the planning phase in Chapter 1:

Research Objectives	Description
1. To analyse the deep learning technique (including CNNs, LSTMs, Transformer and hybrid approach) for emotion recognition.	- The first objective was achieved through a detailed analysis of various deep learning approaches applied in the field of Speech Emotion Recognition (SER) in Chapter 2 Literature Review. A comprehensive literature review of CNNs, LSTMs, Transformers, and hybrid techniques was conducted. Based on these review and analysis on comparison of previous studies, a hybrid SER model which is CNN LSTM being proposed to be implemented as this hybrid model has achieved a great accuracy while been tested datasets like IEMOCAP, Berlin EMO-DB, RAVDESS and SAVEE.
2. To develop an acoustic model that able to recognize emotions (happiness, sadness, anger, etc.) from adult speech using the CREMA-D dataset.	- The second objective was achieved by developing and training the proposed acoustic models which is 2D CNN-LSTM in Chapter 4 Implementation and Testing. Additionally, a baseline experiment was conducted using only a 2D CNN model to serve as a comparative

	reference. This baseline allows for the assessment of the effectiveness of incorporating temporal modelling components, specifically Long Short-Term Memory (LSTM) layers. Both models were trained and tested on the CREMA-D dataset, which consists of adult speech samples labelled with six emotional categories: happiness, sadness, anger, fear, disgust, and neutral. The models have also employed various techniques such as Mel Spectrogram feature extraction and data augmentation to improve generalization.
3. To test the performance of the acoustic model using key metrics like accuracy, precision, recall, and F1-score, and benchmark them against existing studies.	- The third objective was achieved by evaluating both models using key performance metrics, including accuracy, precision, recall, and F1-score in Chapter 5 Result and Analysis. A detailed confusion matrix and classification report were generated to analyse class-wise performance comprehensively. The proposed CNN-LSTM model achieved an overall accuracy of 59.1%, demonstrating improved emotion recognition compared to the baseline CNN model. Additionally, a benchmark study

	was performed to compare the proposed model's performance with existing approaches that utilized different architectures and datasets. This comparison highlighted both the strengths and limitations of the proposed model and offered valuable insights into the impact of dataset characteristics on model performance in Speech Emotion Recognition tasks.

Table 22: The Summary of Research Objectives.

6.3 Research Limitations

Primary limitation encountered in this research lies in the dataset complexity and emotion class distribution within the CREMA-D which has significantly impacted the model's accuracy and performance generalization. While the CREMA-D dataset provides a diverse set of emotional speech samples, it presented several critical issues that hindered the accuracy of the proposed 2D CNN–LSTM model.

Firstly, there exists a high imbalance in speech intensity levels, where over 80% of the audio samples across both the training and testing sets are labelled as "unspecified" intensity. This uneven distribution leaves only a small portion, approximately around 6 to 7% each, of samples representing low, medium, and high intensity speech. As a result, the model becomes biased toward the dominant unspecified intensity and fails to effectively learn prosodic features such as amplitude and pitch that are important in distinguishing between high and low-arousal emotions. Low-arousal emotions like Disgust and Sadness are often expressed subtly and are acoustically similar which making them prone to confusion when the dataset does not clearly differentiate intensity levels.

Moreover, CREMA-D is frequently reported in the literature as one of the most challenging datasets in Speech Emotion Recognition (SER). According to study by Savla et al. (2023), the proposed CNN–LSTM model achieved 83.1% accuracy on RAVDESS dropped to 63.8% on CREMA-D. Similarly, Rafik (2024) reported only 54.1% accuracy on CREMA-D using their Model-B, even though they achieved 98.9% on the TESS dataset. These findings underscore the inherent difficulty posed by CREMA-D's data distribution and complexity, which might affect the performance of SER models, including the proposed model in this study.

Apart from that, the complexity of the CNN-LSTM model architecture itself presents a limitation. The use of deeper convolutional layers with increased numbers of filters or layers

will also increases the number of trainable parameters. This complexity demands greater computational resources and longer training times which may limit the feasibility of extensive hyperparameter tuning or experimentation with larger batch sizes. Consequently, the model's training process becomes more resource-intensive and time-consuming.

6.4 Future Work

Based on the limitations encountered in this research, future research in Speech Emotion Recognition (SER) should focus on both methodological improvements and practical applications, especially in mental health domain.

One potential approach to improve model generalization and address CREMA-D's limitations is the use of advanced data augmentation techniques. SER datasets sometimes suffer from class imbalance and limited diversity especially for subtle or less frequently expressed emotions such as Disgust and Fear. For example, advanced methods like Generative Adversarial Networks (GANs) could be considered for future implementation. Introduced by Goodfellow et al. (2014), GANs are a deep learning architecture consisting of two neural networks: a generator and a discriminator. The generator creates synthetic data samples that aim to mimic real data, while the discriminator evaluates whether a given sample is real or generated.

In SER context, by leveraging an adversarial training process between a generator and a discriminator, GANs can generate realistic and diverse synthetic speech samples that augment underrepresented emotional classes in imbalanced datasets like CREMA-D. Moreover, GANs can also effectively simulate natural variations in pitch, energy, and spectral features which help in enriching the training data with emotion-rich speech (Chatziagapi et al., 2019). These synthetic samples not only help balance the dataset but also introduce greater prosodic and acoustic variation. This will enable SER models to become more sensitive to subtle emotional cues. Apart from that, recent studies by Pappagari et al. (2021) demonstrated that the CopyPaste augmentation method which combining different augmentation techniques such as Neutral CopyPaste (N-CP) and Same Emotion CopyPaste (SE-CP) could help in further diversifies emotional speech datasets and enhances model performance and generalization.

Another important area for future improvement is through the adoption of cross-corpus or cross-validation techniques. Relying on a single dataset like CREMA-D can limit a model's ability to generalize across different speakers, languages, and recording conditions, which is a common challenge in speech emotion recognition (SER). Cross-corpus validation addresses this by training a model on one dataset and testing it on others, such as RAVDESS or TESS, exposing the model to a wider variety of emotional expressions and acoustic conditions. This approach helps mitigate dataset-specific biases and improves the generalization of SER models in real-world scenarios where data variability is high.

Meanwhile, cross-validation is a widely used technique for reliably evaluating model performance within a dataset. It involves dividing the data into *k* equal parts, or folds. The model is trained on *k-1* folds and tested on the remaining fold, repeating this process *k* times so that each fold serves as the test set once. This approach maximizes the use of limited data and helps reduce the risk of overfitting by validating the model on different subsets of the data.

Besides the data-focused strategies, future work should also explore advanced deep learning architectures. Transformer-based models, including those using self-attention mechanisms or architectures like Wav2Vec 2.0, have demonstrated strong capabilities in capturing long-range dependencies and contextual information in speech. (Patamia et al., 2021). Unlike traditional recurrent neural network, transformers use multi-head self-attention to focus on important parts of the speech signal simultaneously which allowing them to understand complex temporal patterns over long sequences (Vaswani et al., 2017). This ability is particularly important for accurately identifying subtle emotional cues in audio signals as emotions often depend on both short-term changes like pitch and long-term context such as emotional tone evolution.

Speech Emotion Recognition (SER) has offered great potential to help mental health care by detecting emotions early, monitoring people continuously, and supporting therapy (Hashem et al., 2023; Olawade et al., 2024). SER systems that have great accuracy and performance generalization can track changes in a person's emotional state by analysing their voice which can help spot early signs of mental health issues like depression, anxiety, or post-traumatic stress disorder. For example, SER can recognize when someone shows long-lasting low-level sadness linked to depression or increased tension in their voice that may indicate anxiety. Moreover, integrating SER into telehealth platforms, empathetic chatbots, or digital therapeutic tools can provide real-time emotional feedback, support timely interventions, and assist clinicians in monitoring patient progress more objectively (Pucci et al., 2023). These applications are particularly valuable for individuals who may struggle to express their emotions verbally or require continuous monitoring outside clinical settings.

In conclusion, advancing research in Speech Emotion Recognition demands a comprehensive and integrated approach. These efforts are essential not only for achieving technical advancements but also for enabling SER to become a transformative tool in mental health care, supporting clinicians and improving patient outcomes through more empathetic, timely, and objective emotional assessment.

6.5 Summary

This chapter summarizes the key findings and contributions of the study on Speech Emotion Recognition (SER) using the CREMA-D dataset. Limitations were identified, particularly the dominance of unspecified intensity samples, which affected the model's ability to recognize subtle emotional variations. Recommendations for future work include advanced data augmentation, cross-corpus validation, and the adoption of transformer-based architectures to enhance model accuracy and generalization. The chapter also highlights the significant potential of SER in mental health applications.

References

Abdallah, M., An Le Khac, N., Jahromi, H., & Delia Jurcut, A. (2021). A Hybrid CNN-LSTM Based Approach for Anomaly Detection Systems in SDNs. *Proceedings of the 16th International Conference on Availability, Reliability and Security*, 1–7. https://doi.org/10.1145/3465481.3469190

Abdelhamid, A. A., El-Kenawy, E.-S. M., Alotaibi, B., Amer, G. M., Abdelkader, M. Y., Ibrahim, A., & Eid, M. M. (2022). Robust Speech Emotion Recognition Using CNN+LSTM Based on Stochastic Fractal Search Optimization Algorithm. *IEEE Access*, *10*, 49265–49284. https://doi.org/10.1109/ACCESS.2022.3172954

Ajibola Alim, S., & Khair Alang Rashid, N. (2018). Some Commonly Used Speech Feature Extraction Algorithms. In R. Lopez-Ruiz (Ed.), *From Natural to Artificial Intelligence—Algorithms and Applications*. IntechOpen. https://doi.org/10.5772/intechopen.80419

Ambartsoumian, A., & Popowich, F. (2018). *Self-Attention: A Better Building Block for Sentiment Analysis Neural Network Classifiers*. https://doi.org/10.18653/v1/P17

Amjad, A., Khan, L., & Chang, H.-T. (2022). Data augmentation and deep neural networks for the classification of Pakistani racial speakers recognition. *PeerJ Computer Science*, *8*, e1053. https://doi.org/10.7717/peerj-cs.1053

Atmaja, B. T., & Sasou, A. (2022). Effects of Data Augmentations on Speech Emotion Recognition. *Sensors*, *22*(16), 5941. https://doi.org/10.3390/s22165941

Ayadi, S. (2024). Speech Emotion Recognition using Dual-Conv2D architecture. *PRZEGLĄD ELEKTROTECHNICZNY*, *1*(6), 211–213. https://doi.org/10.15199/48.2024.06.43

Bahdanau, D., Cho, K., & Bengio, Y. (2014). *Neural Machine Translation by Jointly Learning to Align and Translate*.

Bai, Y. (2022). RELU-Function and Derived Function Review. *SHS Web of Conferences*, *144*, 02006. https://doi.org/10.1051/shsconf/202214402006

Begazo, R., Aguilera, A., Dongo, I., & Cardinale, Y. (2024). A Combined CNN Architecture for Speech Emotion Recognition. *Sensors*, *24*(17), 5797. https://doi.org/10.3390/s24175797

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. *Interspeech 2005*, 1517–1520. https://doi.org/10.21437/Interspeech.2005-446

Cao, H., Beňuš, Š., Gur, R. C., Verma, R., & Nenkova, A. (2014). Prosodic cues for emotion: Analysis with discrete characterization of intonation. *Speech Prosody 2014*, 130–134. https://doi.org/10.21437/SpeechProsody.2014-14

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset. *IEEE Transactions on Affective Computing*, *5*(4), 377–390. https://doi.org/10.1109/TAFFC.2014.2336244

Chatziagapi, A., Paraskevopoulos, G., Sgouropoulos, D., Pantazopoulos, G., Nikandrou, M., Giannakopoulos, T., Katsamanis, A., Potamianos, A., & Narayanan, S. (2019). Data Augmentation Using GANs for Speech Emotion Recognition. *Interspeech 2019*, 171–175. https://doi.org/10.21437/Interspeech.2019-2561

Chauhan, K., Sharma, K. K., & Varma, T. (2021). Speech Emotion Recognition Using Convolution Neural Networks. *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, 1176–1181. https://doi.org/10.1109/ICAIS50930.2021.9395844

Chen, P.-C., Tsai, H., Bhojanapalli, S., Chung, H. W., Chang, Y.-W., & Ferng, C.-S. (2021). A Simple and Effective Positional Encoding for Transformers. *Proceedings of the 2021*

Conference on Empirical Methods in Natural Language Processing, 2974–2988. https://doi.org/10.18653/v1/2021.emnlp-main.236

Chu, H.-C., Zhang, Y.-L., & Chiang, H.-C. (2023). A CNN Sound Classification Mechanism Using Data Augmentation. *Sensors*, *23*(15), 6972. https://doi.org/10.3390/s23156972

Costa, W., Talavera, E., Oliveira, R., Figueiredo, L., Teixeira, J. M., Lima, J. P., & Teichrieb, V. (2023). A Survey on Datasets for Emotion Recognition from Vision: Limitations and In-the-Wild Applicability. *Applied Sciences*, *13*(9), 5697. https://doi.org/10.3390/app13095697

De Lope, J., & Graña, M. (2023). An ongoing review of speech emotion recognition. *Neurocomputing*, *528*, 1–11. https://doi.org/10.1016/j.neucom.2023.01.002

Ding, N., Sethu, V., Epps, J., & Ambikairajah, E. (2012). Speaker variability in emotion recognition—An adaptation based approach. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5101–5104. https://doi.org/10.1109/ICASSP.2012.6289068

Donuk, K. (2022). CREMA-D: Improving Accuracy with BPSO-Based Feature Selection for Emotion Recognition Using Speech. *Journal of Soft Computing and Artificial Intelligence*, *3*(2), 51–57. https://doi.org/10.55195/jscai.1214312

Dutt, A., & Gader, P. (2023). Wavelet Multiresolution Analysis Based Speech Emotion Recognition System Using 1D CNN LSTM Networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *31*, 2043–2054. https://doi.org/10.1109/TASLP.2023.3277291

Ekberg, M., Stavrinos, G., Andin, J., Stenfelt, S., & Dahlström, Ö. (2023). Acoustic Features Distinguishing Emotions in Swedish Speech. *Journal of Voice*, S0892199723001030. https://doi.org/10.1016/j.jvoice.2023.03.010

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, *44*(3), 572–587. https://doi.org/10.1016/j.patcog.2010.09.020

Goh, K. W., Surono, S., Afiatin, M. Y. F., Mahmudah, K. R., Irsalinda, N., Chaimanee, M., & Onn, C. W. (2024). Comparison of Activation Functions in Convolutional Neural Network for Poisson Noisy Image Classification. *Emerging Science Journal*, *8*(2), 592–602. https://doi.org/10.28991/ESJ-2024-08-02-014

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). *Generative Adversarial Networks* (No. arXiv:1406.2661). arXiv. https://doi.org/10.48550/arXiv.1406.2661

Hareli, S., Kafetsios, K., & Hess, U. (2015). A cross-cultural study on emotion expression and the learning of social norms. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01501

Hashem, A., Arif, M., & Alghamdi, M. (2023). Speech emotion recognition approaches: A systematic review. *Speech Communication*, *154*, 102974. https://doi.org/10.1016/j.specom.2023.102974

Hasija, T., Kadyan, V., Guleria, K., Alharbi, A., Alyami, H., & Goyal, N. (2022). Prosodic Feature-Based Discriminatively Trained Low Resource Speech Recognition System. *Sustainability*, *14*(2), 614. https://doi.org/10.3390/su14020614

Hu, W., & Thing, V. L. L. (2024). *CPE-Identifier: Automated CPE identification and CVE summaries annotation with Deep Learning and NLP* (No. arXiv:2405.13568). arXiv. https://doi.org/10.48550/arXiv.2405.13568

Ibrahim, N. J., Idna Idris, M. Y., Mohd Yusoff, M. Y. @ Z., Abdul Rahman, N. N., & Izzi Dien, M. (2019). ROBUST FEATURE EXTRACTION BASED ON SPECTRAL AND

PROSODIC FEATURES FOR CLASSICAL ARABIC ACCENTS RECOGNITION. *Malaysian Journal of Computer Science*, 46–72. https://doi.org/10.22452/mjcs.sp2019no3.4

Ioffe, S., & Szegedy, C. (2015). *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift* (No. arXiv:1502.03167). arXiv. https://doi.org/10.48550/arXiv.1502.03167

Islam, M. M. M., Kabir, M. A., Sheikh, A., Saiduzzaman, M., Hafid, A., & Abdullah, S. (2024). Enhancing Speech Emotion Recognition Using Deep Convolutional Neural Networks. *2024 9th International Conference on Machine Learning Technologies (ICMLT)*, 95–100. https://doi.org/10.1145/3674029.3674045

Izard, C. E. (2009). Emotion Theory and Research: Highlights, Unanswered Questions, and Emerging Issues. *Annual Review of Psychology*, *60*(1), 1–25. https://doi.org/10.1146/annurev.psych.60.110707.163539

Jackson, P. J. B., & Haq, S. U. (2011). Surrey Audio-Visual Expressed Emotion (SAVEE) database. *University of Surrey*.

Kamble, V. V., Deshmukh, R. R., Karwankar, A. R., Ratnaparkhe, V. R., & Annadate, S. A. (2015). Emotion Recognition for Instantaneous Marathi Spoken Words. In S. C. Satapathy, B. N. Biswal, S. K. Udgata, & J. K. Mandal (Eds.), *Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014* (Vol. 328, pp. 335–346). Springer International Publishing. https://doi.org/10.1007/978-3-319-12012-6_37

Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. (2019). Speech Emotion Recognition Using Deep Learning Techniques: A Review. *IEEE Access*, *7*, 117327–117345. https://doi.org/10.1109/ACCESS.2019.2936124

Kurniawan, F., Sulaiman, S., Konate, S., & Abdalla, M. A. A. (2023). Deep learning approaches for MIMO time-series analysis. *International Journal of Advances in Intelligent Informatics*, *9*(2), 286. https://doi.org/10.26555/ijain.v9i2.1092

Labied, M., & Belangour, A. (2021). Automatic Speech Recognition Features Extraction Techniques: A Multi-criteria Comparison. *International Journal of Advanced Computer Science and Applications*, *12*(8). https://doi.org/10.14569/IJACSA.2021.0120821

Li, Y., Baidoo, C., Cai, T., & Kusi, G. A. (2019). Speech Emotion Recognition Using 1D CNN with No Attention. *2019 23rd International Computer Science and Engineering Conference (ICSEC)*, 351–356. https://doi.org/10.1109/ICSEC47112.2019.8974716

Liu, G., Cai, S., & Wang, C. (2023). Speech emotion recognition based on emotion perception. *EURASIP Journal on Audio, Speech, and Music Processing*, *2023*(1), 22. https://doi.org/10.1186/s13636-023-00289-4

Liu, X., Shi, T., Zhou, G., Liu, M., Yin, Z., Yin, L., & Zheng, W. (2023). Emotion classification for short texts: An improved multi-label method. *Humanities and Social Sciences Communications*, *10*(1), 306. https://doi.org/10.1057/s41599-023-01816-6

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE*, *13*(5), e0196391. https://doi.org/10.1371/journal.pone.0196391

Lu, Y., & Salem, F. M. (2017). Simplified gating in long short-term memory (LSTM) recurrent neural networks. *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, 1601–1604. https://doi.org/10.1109/MWSCAS.2017.8053244

Malashin, I., Tynchenko, V., Gantimurov, A., Nelyub, V., & Borodulin, A. (2024). *Applications of Long Short-Term Memory (LSTM) Networks in Polymeric Sciences: A Review*.

Mary, L., & Yegnanarayana, B. (2008). Extraction and representation of prosodic features for language and speaker recognition. *Speech Communication*, *50*(10), 782–796. https://doi.org/10.1016/j.specom.2008.04.010

Nanni, L., Maguolo, G., Brahnam, S., & Paci, M. (2021). An Ensemble of Convolutional Neural Networks for Audio Classification. *Applied Sciences*, *11*(13), 5796. https://doi.org/10.3390/app11135796

Narimisaei, J., Naeim, M., Imannezhad, S., Samian, P., & Sobhani, M. (2024). Exploring emotional intelligence in artificial intelligence systems: A comprehensive analysis of emotion recognition and response mechanisms. *Annals of Medicine & Surgery*. https://doi.org/10.1097/MS9.0000000000002315

Nogueiras, A., Moreno, A., Bonafonte, A., & Mariño, J. B. (2001). Speech emotion recognition using hidden Markov models. *7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, 2679–2682. https://doi.org/10.21437/Eurospeech.2001-627

Olawade, D. B., Wada, O. Z., Odetayo, A., David-Olawade, A. C., Asaolu, F., & Eberhardt, J. (2024). Enhancing mental health with Artificial Intelligence: Current trends and future prospects. *Journal of Medicine, Surgery, and Public Health*, *3*, 100099. https://doi.org/10.1016/j.glmedi.2024.100099

Pan, S.-T., & Wu, H.-J. (2023). Performance Improvement of Speech Emotion Recognition Systems by Combining 1D CNN and LSTM with Data Augmentation. *Electronics*, *12*(11), 2436. https://doi.org/10.3390/electronics12112436

Pappagari, R., Villalba, J., Zelasko, P., Moro-Velazquez, L., & Dehak, N. (2021). CopyPaste: An Augmentation Method for Speech Emotion Recognition. *ICASSP 2021 - 2021 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6324–6328. https://doi.org/10.1109/ICASSP39728.2021.9415077

Patamia, R. A., Jin, W., Acheampong, K. N., Sarpong, K., & Tenagyei, E. K. (2021). Transformer Based Multimodal Speech Emotion Recognition with Improved Neural Networks. *2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML)*, 195–203. https://doi.org/10.1109/PRML52754.2021.9520692

Patil, K. J., Zope, P. H., & Suralkar, S. R. (2012). Emotion Detection From Speech Using Mfcc & Gmm. *International Journal of Engineering Research*, *1*(9).

Pei, G., Li, H., Lu, Y., Wang, Y., Hua, S., & Li, T. (2024). Affective Computing: Recent Advances, Challenges, and Future Trends. *Intelligent Computing*, *3*, 0076. https://doi.org/10.34133/icomputing.0076

Pichora-Fuller, M. K., & Dupuis, K. (2020). Toronto emotional speech set (TESS). *Borealis*. https://doi.org/10.5683/SP2/E8H2MF

Pucci, F., Fedele, P., & Dimitri, G. M. (2023). Speech emotion recognition with artificial intelligence for contact tracing in the COVID-19 pandemic. *Cognitive Computation and Systems*, *5*(1), 71–85. https://doi.org/10.1049/ccs2.12076

Rafik, H. D. (2024). *Detection and classification of Emotion Recognition System for TESS and Crema-d Audio Datasets Using Hybrid Deep Learning Architecture*.

Rezapour Mashhadi, M. M., & Osei-Bonsu, K. (2023). Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest. *PLOS ONE*, *18*(11), e0291500. https://doi.org/10.1371/journal.pone.0291500

Sajun, A. R., Zualkernan, I., & Sankalpa, D. (2024). A Historical Survey of Advances in Transformer Architectures. *Applied Sciences*, *14*(10), 4316. https://doi.org/10.3390/app14104316

Sak, H., Senior, A., & Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. *Interspeech 2014*, 338–342. https://doi.org/10.21437/Interspeech.2014-80

Salian, B., Narvade, O., Tambewagh, R., & Bharne, S. (2021). Speech Emotion Recognition using Time Distributed CNN and LSTM. *ITM Web of Conferences*, *40*, 03006. https://doi.org/10.1051/itmconf/20214003006

Savla, M., Gopani, D., Ghuge, M., Chaudhari, S., & Raundale, P. (2023). Sentiment Analysis of Human Speech using Deep Learning. *2023 3rd International Conference on Intelligent Technologies (CONIT)*, 1–6. https://doi.org/10.1109/CONIT59222.2023.10205915

Scherer, K. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1–2), 227–256. https://doi.org/10.1016/S0167-6393(02)00084-5

Schuller, B., Rigoll, G., & Lang, M. (2003). *HIDDEN MARKOV MODEL-BASED SPEECH EMOTION RECOGNITION*.

Sefara, T. J. (2019). The Effects of Normalisation Methods on Speech Emotion Recognition. *2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, 1–8. https://doi.org/10.1109/IMITEC45504.2019.9015895

Sharan, R. V., Mascolo, C., & Schuller, B. W. (2024). Emotion Recognition from Speech Signals by Mel-Spectrogram and a CNN-RNN. *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 1–4. https://doi.org/10.1109/EMBC53108.2024.10782952

Sharma, S., & Mehra, R. (2019). Implications of Pooling Strategies in Convolutional Neural Networks: A Deep Insight. *Foundations of Computing and Decision Sciences*, *44*(3), 303–330. https://doi.org/10.2478/fcds-2019-0016

Singh, J., Saheer, L. B., & Faust, O. (2023). Speech Emotion Recognition Using Attention Model. *International Journal of Environmental Research and Public Health*, *20*(6), 5140. https://doi.org/10.3390/ijerph20065140

Sonkar, S., & Baraniuk, R. G. (2023). *Investigating the Role of Feed-Forward Networks in Transformers Using Parallel Attention and Feed-Forward Net Design* (No. arXiv:2305.13297). arXiv. http://arxiv.org/abs/2305.13297

Srinivas, D. P. V. S., Akshaya, A., Singh, B. P., & Mounika, D. (2024). *Recognising Speech Based Emotion Using CNN in Deep Learning*. *45*(3).

Stanley, E., DeMattos, E., Klementiev, A., Ozimek, P., Clarke, G., Berger, M., & Palaz, D. (2023). Emotion Label Encoding Using Word Embeddings for Speech Emotion Recognition. *INTERSPEECH 2023*, 2418–2422. https://doi.org/10.21437/Interspeech.2023-1591

Swain, T., Anand, U., Aryan, Y., Khanra, S., Raj, A., & Patnaik, S. (2021). Performance Comparison of LSTM Models for SER. *In Proceedings of International Conference on Communication, Circuits, and Systems*, *IC3S 2020*, 427–433.

Tamati, T. N., Gilbert, J. L., & Pisoni, D. B. (2013). Some Factors Underlying Individual Differences in Speech Recognition on PRESTO: A First Report. *Journal of the American Academy of Audiology*, *24*(07), 616–634. https://doi.org/10.3766/jaaa.24.7.10

Tao, H., Shan, S., Hu, Z., Zhu, C., & Ge, H. (2022). Strong Generalized Speech Emotion Recognition Based on Effective Data Augmentation. *Entropy*, *25*(1), 68. https://doi.org/10.3390/e25010068

Tracey, B., Volfson, D., Glass, J., Haulcy, R., Kostrzebski, M., Adams, J., Kangarloo, T., Brodtmann, A., Dorsey, E. R., & Vogel, A. (2023). Towards interpretable speech biomarkers: Exploring MFCCs. *Scientific Reports*, *13*(1), 22787. https://doi.org/10.1038/s41598-023-49352-2

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is All you Need*.

Voita, E., Talbot, D., Moiseev, F., Sennrich, R., & Titov, I. (2019). Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5797–5808. https://doi.org/10.18653/v1/P19-1580

Wall, D., & Horák, T. (2007). Using baseline studies in the investigation of test impact. *Assessment in Education: Principles, Policy & Practice*, *14*(1), 99–116. https://doi.org/10.1080/09695940701272922

Wolf-Monheim, F. (2024). *Spectral and Rhythm Features for Audio Classification with Deep Convolutional Neural Networks* (No. arXiv:2410.06927). arXiv. http://arxiv.org/abs/2410.06927

Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: An overview and application in radiology. *Insights into Imaging*, *9*(4), 611–629. https://doi.org/10.1007/s13244-018-0639-9

Yang, X., Yu, S., & Xu, W. (2024). *Improvement and Implementation of a Speech Emotion Recognition Model Based on Dual-Layer LSTM* (No. arXiv:2411.09189). arXiv. https://doi.org/10.48550/arXiv.2411.09189

Younis, E. M. G., Mohsen, S., Houssein, E. H., & Ibrahim, O. A. S. (2024). Machine learning for human emotion recognition: A comprehensive review. *Neural Computing and Applications*, *36*(16), 8901–8947. https://doi.org/10.1007/s00521-024-09426-2

Zhang, S., & Pell, M. D. (2022). Cultural differences in vocal expression analysis: Effects of task, language, and stimulus-related factors. *PLOS ONE*, *17*(10), e0275915. https://doi.org/10.1371/journal.pone.0275915

Zhao, J., Mao, X., & Chen, L. (2018). Learning deep features to recognise speech emotion using merged deep CNN. *IET Signal Processing*, *12*(6), 713–721. https://doi.org/10.1049/iet-spr.2017.0320

Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control*, *47*, 312–323. https://doi.org/10.1016/j.bspc.2018.08.035

Zhao, Y., & Shu, X. (2023). Speech emotion analysis using convolutional neural network (CNN) and gamma classifier-based error correcting output codes (ECOC). *Scientific Reports*, *13*(1), 20398. https://doi.org/10.1038/s41598-023-47118-4

Zhou, Z., Yuan, H., & Cai, X. (2023). Rock Thin Section Image Identification Based on Convolutional Neural Networks of Adaptive and Second-Order Pooling Methods. *Mathematics*, *11*(5), 1245. https://doi.org/10.3390/math11051245