DEEP HIERARCHICAL WEIGHTED LATE FUSION MODEL FOR MULTIMODAL EMOTION RECOGNITION

By

NURUL NADHRAH BINTI KAMARUZAMAN

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Doctor of Philosophy

August 2024

FSKTM 2024 23

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

DEEP HIERARCHICAL WEIGHTED LATE FUSION MODEL FOR MULTIMODAL EMOTION RECOGNITION

By

NURUL NADHRAH BINTI KAMARUZAMAN

August 2024

Chairman : Nor Azura binti Husin, PhD
Faculty : Computer Science and Information Technology

Emotion recognition gains its relevance in diverse areas such as healthcare, education, business marketing and entertainment. As time goes on, research in emotion recognition analysis has developed from the traditional single modal to complex multimodal analysis. Extraction of data from various modalities such as audio, text and video are required in order to gain a meaningful pattern, thus increasing the accuracy of the recognition. In this study, the focus was given on implementing the late fusion approach to integrate three modalities for multimodal emotion recognition. Nevertheless, the late fusion approach faces several challenges, including the loss of fine-grained temporal and spatial information, class imbalance across modalities and reliance on individual model training performance. To address these issues, this study presents the Deep Hierarchical Weighted Late Fusion (DHWLF) model, which is designed for an accurate multimodal emotion recognition. DHWLF employs deep hierarchical weighted late fusion method to effectively reduce the loss of fine-grained temporal and spatial information. This study also introduces individual emotion recognition models for audio, text, and video image modalities using deep learning

techniques. Specifically, for the audio, this study introduces the SMOTE-2DCNN model, which combines the Synthetic Minority Oversampling Technique (SMOTE) method with a 2-Dimensional Convolutional Neural Network (2DCNN) to handle imbalanced class distribution and classify emotions accurately. For the text, this study implements the Bidirectional Encoder Representations from Transformers (BERT) language model known as TextBERT model to capture semantic and contextual information from textual data. Lastly, for the video, this study proposes the Fer2DCNN model, which utilizes the Cascade Haar Classifier to detect facial regions before feeding them into a 2DCNN for emotion classification. To evaluate the effectiveness of the DHWLF model, extensive experiments were conducted. The results demonstrated its superiority with the DHWLF model achieving 90% accuracy, outperforming both individual emotion recognition models and other state-of-the-art multimodal emotion recognition models from prior studies.

Keywords: Late Fusion, Weighted Late Fusion, Multimodal, Emotion Recognition, Deep Learning

SDG: GOAL 3: Good Health and Well-Being, GOAL 4: Quality Education, GOAL 9: Industry, Innovation and Infrastructure Quality

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

DEEP HIERARCHICAL WEIGHTED LATE FUSION MODEL FOR MULTIMODAL EMOTION RECOGNITION

By

NURUL NADHRAH BINTI KAMARUZAMAN

Ogos 2024

Pengerusi : Nor Azura binti Husin, PhD
Fakulti : Sains Komputer dan Teknologi Maklumat

Pengecaman emosi mempunyai peranan yang penting dalam pelbagai bidang seperti penjagaan kesihatan, pendidikan, peniagaan pemasaran dan hiburan. Seiring berjalannya waktu, penyelidikan dalam analisis pengecaman emosi telah berkembang daripada analisis modal tunggal tradisional kepada analisis multimodal yang kompleks. Pengekstrakan data daripada pelbagai modaliti seperti audio, teks dan video diperlukan untuk mendapatkan corak yang bermakna, sekali gus meningkatkan ketepatan dalam pengecaman. Dalam kajian ini tumpuan diberikan terutamanya kepada penggunaan pendekatan gabungan lewat untuk mengintegrasikan tiga modaliti bagi pembangunan modal pengecaman emosi pelbagai mod. Walau bagaimanapun, pendekatan gabungan lewat menghadapi cabaran termasuk kehilangan maklumat temporal dan spatial yang terperinci, ketidakseimbangan data antara modaliti, dan pergantungan pada prestasi latihan model individu. Untuk menangani isu-isu ini, kajian ini memperkenalkan model DHWLF yang direka untuk tugas pengecaman emosi yang tepat, yang melibatkan pelbagai modaliti. DHWLF memanfaatkan kaedah gabungan lewat berwajaran hierarki untuk mengurangkan kehilangan maklumat

temporal dan spatial yang terperinci. Tambahan pula, penyelidikan ini memperkenalkan model pengecaman emosi individu untuk modaliti imej audio, teks dan video menggunakan teknik pembelajaran mendalam. Secara khusus, untuk audio, kajian ini memperkenalkan model SMOTE-2DCNN, yang menggabungkan kaedah SMOTE dengan rangkaian neural terkawal konvolusi 2 dimensi (2DCNN) untuk mengendalikan taburan data yang tidak seimbang dan mengklasifikasikan kelas emosi dengan tepat. Untuk teks, kajian ini menggunakan model bahasa BERT yang dekenali sebadai TexTBERT model untuk menangkap maklumat semantik dan kontekstual dari data teks. Akhir sekali, untuk video, kajian ini mencadangkan model Fer2DCNN, yang menggunakan pengkelas Haar berkaskad untuk untuk mengesan kawasan muka sebelum memasukkannya ke dalam 2DCNN untuk pengelasan emosi. Untuk menilai keberkesanan, eksperimen yang meluas telah dijalanka. Keputusan menunjukkan keunggulannya dengan model DHWLF mencapai ketepatan 90%, mengatasi kedua-dua model pengecaman emosi individu dan model pengecaman emosi multimodal terkini yang lain daripada kajian terdahulu.

Kata Kunci: Gabungan Lewat, Gabungan Lewat Berwajaran, Mod Berbilang, Pengecaman Emosi, *Deep Learning*

SDG: MATLAMAT 3: Good Health and Well-Being, MATLAMAT 4: Quality Education, MATLAMAT 9: Industry, Innovation and Infrastructure Quality

ACKNOWLEDGEMENTS

In the name of Allah, the Most Gracious, the Most Merciful.

With a heart overflowing with gratitude, I begin this acknowledgment by thanking Allah, the Almighty, for bestowing upon me the strength, patience and guidance to traverse this challenging and fulfilling journey of completing my thesis.

I would like to express my sincere appreciation to my supervisor, Dr. Nor Azura Husin and my committee members, Assoc. Professor Datin Dr. Norwati Mustapha, Assoc. Professor Dr. Razali Yaakob and Prof. Dr. Firdaus Binti Mamat @ Mukhtar for their invaluable guidance and constructive feedback throughout this research endeavor. Your expertise, wisdom, and relentless dedication have shaped and polished my work, enabling me to refine my ideas and deepen my understanding. I am immensely grateful for the knowledge and skills I have acquired under your mentorship.

A special thanks goes to the distinguished lecturers at Universiti Putra Malaysia (UPM), especially those from the Faculty of Computer Science and Information Technology (FCSIT). Their vast knowledge, dedication to teaching, and unwavering commitment to excellence have nurtured my intellectual growth and broadened my horizons. Their passion for imparting knowledge has been a constant source of inspiration throughout my academic journey.

To my beloved family, especially my mom, dad, husband and sisters, words fail to express the depth of my gratitude. Your unconditional love, unwavering support and endless encouragement have been the bedrock of my journey. Your prayers have

provided me with the strength to persevere and pursue my dreams. All your sacrifices and selflessness have been the pillars of my success and I am eternally grateful for your boundless love and unwavering faith in my abilities.

I would also like to extend my heartfelt appreciation to my friends and all those who have directly and indirectly contributed to the completion of this thesis. Their encouragement, moral support, and willingness to lend a helping hand, whether through discussions, brainstorming sessions or providing valuable resources, have been invaluable. Their belief in my potential and their unwavering support throughout this arduous journey have been instrumental in shaping my ideas and providing much-needed motivation during moments of doubt

Lastly, I express my profound gratitude to all the individuals who have contributed to my academic and personal growth, even if our paths have crossed only briefly. Your assistance, guidance, and encouragement have had a profound impact on my journey, and I am deeply indebted to each and every one of you.

This thesis would not have been possible without the support, encouragement, and assistance of all those mentioned above. I am humbled by the collective efforts of these remarkable individuals who have believed in me and accompanied me on this challenging yet rewarding endeavor. Thank you all from the bottom of my heart.

May Allah bless you abundantly and grant you the highest place in Jannah.

Nurul Nadhrah Binti Kamaruzaman, 2024

This thesis was submitted to the Senate of the Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee were as follows:

Nor Azura binti Husin, PhD
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Norwati binti Mustapha, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

Razali bin Yaakob, **PhD**
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

Firdaus binti Mamat @ Mukhtar, PhD
Professor
Faculty of Medicine and Health Sciences
Universiti Putra Malaysia
(Member)

ZALILAH MOHD SHARIFF, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 17 April 2025

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

2DCNN	Two-Dimensional Convolutional Neural Network
3DCNN	Three-Dimensional Convolutional Neural Network
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area under the ROC Curve
BERT	Bidirectional Encoder Representations from Transformers
BRNN	Bidirectional Recurrent Neural Network
CNN	Convolutional Neural Network
DBN	Deep Belief Network
DFT	Discrete Fourier Transform
DHWLF	Deep Hierarchical Weighted Late Fusion
DL	Deep Learning
DTW	Dynamic Time Warping
EEG	Electroencephalogram
FER	Facial Expression Recognition
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GFCC	Gammatone Frequency Cepstral Coefficients
GMM	Gaussian Mixture Models
GPU	Graphical Processing Unit
GRNN	Gated Recurrent Neural Networks
GRU	Gated Recurrent Units
HCI	Human Computer Interactions
HMM	Hidden Markov Model

IDE	Integrated development environment
IEMOCAP	Interactive Emotional Dyadic Motion Captur
JAFEE	Japanese Female Facial Expression
LBP	Local Binary Pattern
LFPC	Log-Frequency Power Coefficients
LSTM	Long- Short Term Memory
MFCC	Mel Frequency Cepstral Coefficient
MLM	Masked Language Modeling
MRI	Magnetic Resonance Imaging
NLP	Natural Language Processing
NN	Neural Network
NSP	Next Sentence Prediction
RELU	Rectified Linear Unit
RFC	Random Forest Classifier
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
SAG	Stochastic Average Gradient
SEO	Search Engine Optimization
SER	Speech Emotion Recognition
SIFT	Scale Invariant Feature Transform
SMOTE	Synthetic Minority Oversampling Technique
STFT	Short-Time Fourier-Transform
SVM	Support Vector Machine
TEO	Teager Energy Operator
TER	Text Emotion Recognition
TN	True Negative

TP	True Positive
TPR	True Positive Rate
UA	Unweighted Average
WA	Weighted Average

CHAPTER 1

INTRODUCTION

1.1 Background of Study

Emotions are essential for many daily tasks, including learning, communicating, making decisions, raising awareness in the community, and much more (Poria et al., 2017a). The goal of many researchers has been to give machines more cognitive ability to understand and identify emotions. The potential applications of emotion recognition in healthcare, education, business marketing, and entertainment have made it a prominent research area in recent years. It is now essential to accurately identify and comprehend human emotions from a variety of modalities including text, audio, and video, in order to create intelligent systems that can communicate and interact with people in an efficient manner.

Conventional approaches to recognizing emotions have frequently concentrated on single modalities alone, ignoring the important information that can be gleaned from a variety of data modalities (Z. Huang et al., 2014; Jung et al., 2015; Ziemer & Korkmaz, 2017) . Nonetheless, recent developments in deep learning methodologies have opened the door for the creation of multimodal fusion models, which are capable of efficiently fusing data from several modalities. One advantage of using multiple modalities in emotion recognition is that it reduces the instability of using a single modality, which can improve the accuracy rate of emotion recognition to some extent and lead to improved performance on emotion recognition tasks (Song et al., 2019).

There are two common approaches to integrate these multiple modalities which are early fusion (feature-level fusion) and late fusion (decision-level fusion). Feature-level early fusion involves combining the raw features from different modalities into a single representation before feeding it into the classifier, while decision-level late fusion combines outputs of individual classifiers trained on each modality after they have processed their respective inputs. Despite the fact that feature-level early fusion was used in many earlier studies on multimodal emotion recognition, (Hazarika, Poria, et al., 2018; Majumder et al., 2018; Poria et al., 2018) as it seems intuitive and computationally effective, it frequently encountered difficulties because of the heterogeneity and differing degrees of reliability among modalities.

One of a significant issue with feature-level early fusion is the possibility for feature misalignment between modalities (Pfeuffer & Dietmayer, 2018). Effective combination of features extracted from different modalities can be challenging because of differences in their scales, temporal resolutions, and semantic meanings. For instance, text features may convey the emotion expressed in written language, whereas video features might capture body language and facial expressions. It can be difficult to meaningfully align these dissimilar features, which could lead to noise amplification or information loss.

Decision-level late fusion overcomes these challenges by allowing individual modality-specific classifiers to operate independently and exclusively on their own modalities (Tsanousa et al., 2019). This approach avoids the requirement for feature alignment while preserving the distinctive qualities and information content of each modality. By combining decisions or outputs from multiple classifiers at a higher

level, decision-level late fusion enhances the robustness and flexibility of multimodal information integration, which may lead to a better emotion recognition performance (Tripathi et al., 2019).

Leveraging hierarchical structures in deep learning enhances the benefits of decision-level late fusion. This allows for the integration of multiple levels of abstraction, capturing both low-level features and high-level contextual information (Gu et al., 2018). Such integration facilitates the understanding of complex emotional cues. For example, by combining hierarchical structures such as convolutional neural networks (CNNs) for feature extraction and recurrent neural networks (RNNs) for temporal modelling, the system is able to capture both spatial and temporal dependencies within multimodal data such as video of facial expression, audio and text of speech. Building on previous works that have demonstrated the efficacy of hierarchical deep learning architectures in capturing complex patterns (Jiao et al., 2019; Majumder et al., 2018), decision-level late fusion extends these benefits by allowing for the integration of multiple sources of information at a later stage. This improves overall recognition accuracy and facilitates better fusion of complementary features.

In this study, focus is given on developing a multimodal fusion model for emotion recognition based on deep hierarchical late fusion approach to effectively fuse three different modalities, namely audio, text and video. This fusion strategy enables to exploit the complementary nature of the modalities by sequentially aggregating the predictions of the individual models at different levels of the hierarchy. Each modality is initially trained individually using the deep learning model that is specifically proposed and designed by this study, incorporating state-of-the-art architectures and

techniques. This study also highlights the developing and training process of individuals models for each modality involved. This process is essential for achieving optimal results because it allows for tailored adjustments of hyperparameters and extension experiments to fine-tune each modality prior to the fusion process. By optimizing each model individually, it maximizes each modality's ability to effectively capture its unique features. This approach ensures that the fused model benefits from the best-performing components thus, increase the overall performance and robustness in multimodal emotion recognition tasks.

The study holds the promise of significantly improving the accuracy and robustness of emotion recognition systems. By considering multiple modalities and implementing deep hierarchical late fusion approach, this study aims to capture a more comprehensive representation of human emotions by incorporating cues from audio, text, and video data modalities. This research contributes to the growing body of literature on multimodal fusion techniques and provides valuable insights into the design and development of effective emotion recognition models.

1.2 Problem Statement

The advantage of the decision-level late fusion approach in multimodal emotion recognition is its flexibility to diverse data sources. Due to the possibility that every modality has distinct features and noise characteristics, the late fusion enables the integration of varied data without requiring an extensive pre-processing in order to align features between modalities (Shankar et al., 2022) . This flexibility can lead to improved robustness and adaptability of the emotion recognition system, especially in real-world scenarios where data may be noisy or incomplete.

Despite these advantages, late fusion methods encounter a number of challenges that hinder their effectiveness in accurately recognizing emotions. One of the main issues with the late fusion approach is the loss of fine-grained temporal and spatial information (Cimtay et al., 2020). The loss of fine-grained temporal and spatial information occurs when important details related to time and space in multimodal data are not fully captured during the fusion process. Temporal information pertains to patterns and changes over time, such as the sequence of spoken words or variations in facial expressions in a video. In contrast, spatial information focuses on how features are arranged and interact within a space, like the position and movement of facial landmarks in an image or video. This loss of detailed information is particularly problematic in the late fusion approach, where modalities are usually fused after performing individual processing and training. When multiple modalities capture subtle emotional cues over different timeframes and spatial contexts, the fusion mechanism often fails to preserve the important information, resulting in a diluted representation of emotions and decreased recognition accuracy. For example, consider a scenario where an individual expresses a mix of neutral and sadness. The fusion of facial expressions and speech may fail to capture the nuanced shifts in emotion over time and lead to misinterpretation of the overall emotional state.

Another challenge in late fusion for multimodal emotion recognition is the inherent class imbalance among modalities. The availability and quality of data can differ among modalities, including text, audio, and visual inputs. This can result in an uneven distribution of information during fusion (Meng et al., 2023; Wagner et al., 2011) . For example, in a dataset primarily composed of textual data with fewer audio-visual

samples, the late fusion method may prioritize textual information over visual or auditory cues. This prioritization can lead to a biased representation of emotions.

Apart from that, the late fusion for multimodal emotion recognition encounters significant challenges due to its reliance on prior individual model training performance. Integrating outputs from text, audio, and visual cues at a later stage means the final prediction accuracy depends heavily on each modality's model quality (Pandeya & Lee, 2021; S. Zhang et al., 2024). Moreover, data heterogeneity across modalities complicates alignment and integration which lowers the overall performance (S. Li & Tang, 2024). Thus, ensuring robustness in each modality's model training is essential for achieving accurate and reliable outcomes in late fusion multimodal systems.

In this study, the focus is on addressing the limitations of late fusion methods in multimodal emotion recognition by developing the best rules for determining the most accurate way to fuse all individually trained models (R. Gupta et al., 2016; Y. Huang et al., 2017a; Suen & Lam, 2000), involving three modalities namely audio, text and videos. Building upon previous works, incorporation of data augmentation techniques to overcome class imbalance among modalities is proposed in order to ensure a more equitable distribution of information during fusion (Dablain et al., 2022; Elyan et al., 2021). Furthermore, the focus will be on developing individual models using deep learning methods, with an emphasis on extracting salient features from each modality while also capturing contextual dependencies in emotion recognition tasks (Devlin et al., 2018; J. Li et al., 2020; Mao et al., 2014). This approach involves pre-processing features and hyperparameter tuning to optimize the performance of each individual

model. By employing these strategies in addressing the abovementioned issues, this study aim is to enhance the robustness and accuracy of multimodal fusion emotion recognition systems.

1.3 Research Question

Based on the problem statements stated, the following research questions have been developed to determine the path of this study:

i. How can the best rule be proposed for the most accurate fusion of individually trained models in late fusion methods for multimodal emotion recognition?

ii. What data augmentation techniques can be proposed to address class imbalance among modalities in multimodal emotion recognition?

iii. How can deep learning-based individual models be proposed to extract salient features and capture contextual dependencies in emotion recognition tasks across audio, text, and video modalities?

1.4 Objective

The main objective of this study is to develop an accurate multimodal fusion model for emotion recognition based on three modalities which are audio, text and video.

Three specific objectives to be achieved from this research are:

i. To propose the best rule for the most accurate fusion of individually trained models in late fusion methods for multimodal emotion recognition.

ii. To propose data augmentation techniques in order to address class imbalance among modalities

iii. To propose deep learning-based individual models that capable of extracting salient features and capturing contextual dependencies in emotion recognition tasks

1.5 Connection between Problem Statement, Objective and Research Question

The most important aspect before moving forward with this study is to ensure that the problem statement, objective and research question are in alignment. Table 1-1 presents the connection between the problem statement, objective and research question of this study.

Table 1.1 : Connection between Problem Statement, Objective and Research Question

Problem Statement	Objective	Research Question
Loss of fine-grained temporal and spatial information during late fusion	To propose the best rule for the most accurate fusion of individually trained models in late fusion methods for multimodal emotion recognition.	How can the best rule be proposed for the most accurate fusion of individually trained models in late fusion methods for multimodal emotion recognition?
Class Imbalance among modalities	To propose data augmentation techniques in order to address class imbalance among modalities	What data augmentation techniques can be proposed to address class imbalance among modalities in multimodal emotion recognition?
Reliance on the performance of prior individual model training.	To propose deep learning-based individual models that capable of extracting salient features and capturing contextual dependencies in emotion recognition tasks	How can deep learning-based individual models be proposed to extract salient features and capture contextual dependencies in emotion recognition tasks across audio, text, and video modalities?

1.6 Scope

This study focuses on developing rules for the most accurate late fusion of individually trained models involving three modalities which are audio, text and video. The

IEMOCAP dataset was used to provide multimodal data for emotion recognition, which includes audio, text, and video. Rather than using all of the emotions in the dataset, the study focuses on four emotions which are angry, happy, sad, and neutral. The audio modality in IEMOCAP consists of speech recordings that capture vocal emotion cues such as tone and pitch. The text modality offers transcriptions of the dialogues, providing linguistic content for emotion detection. Meanwhile, the video modality includes visual data of facial expressions and body language, which are essential for identifying non-verbal emotional signals.

This study also proposes data augmentation techniques to overcome class imbalance among modalities. Furthermore, this study involves in designing deep learning-based individual models capable of extracting salient features and capturing contextual dependencies in emotion recognition tasks for audio, text, and video modalities. This includes optimizing hyperparameters and preprocessing techniques for individual models and evaluating the overall robustness and accuracy of the proposed multimodal fusion model.

1.7 Significance of Contribution

This study significantly contributes to the field of multimodal emotion recognition through the specifics of the following contributions:

i. A novel deep hierarchical late fusion model equipped with optimized fusion rules to enhance the integration of information from multiple modalities in emotion recognition tasks.

ii. Introducing data augmentation techniques tailored to handle imbalanced class across modalities to ensure a more balanced distribution of information during fusion.

iii. Development of individual models based on deep learning methods for all three modalities, audio, text, and video, with an emphasis on extracting salient features and capturing contextual dependencies for an accurate emotion recognition across diverse datasets.

1.8 Thesis Organization

This thesis consists of eight chapters with a few sections and subsections in each chapter.

Chapter 1 provides an introduction to the research, beginning with a presentation of the research background, including the identification of the problem statement and the formulation of research questions. The chapter also outlines the objectives of the study, the scope of the research, and the anticipated contributions and significance of the findings. This chapter sets the foundation for the subsequent chapters by establishing the context and rationale for the investigation.

Chapter 2 focuses on the literature review, which delves into the concepts of emotion, deep learning models, and the application of deep learning in emotion recognition. It specifically examines the utilization of multimodal data, encompassing audio, text and video information, in the field of emotion recognition. This chapter synthesizes existing knowledge and establishes the theoretical framework upon which the subsequent chapters build.

Chapter 3 focuses on the research methodology employed in the study. It details the overall approach and research design, outlining the steps taken to conduct the research.

Chapter 4 explores the development of the multimodal fusion classification model for emotion recognition, incorporating audio, text, and video data. It discusses the fusion approach employed in combining the different modalities and elaborates on the algorithmic details of the multimodal model. This chapter presents the experimental setup, along with the results and discussions, examining the effectiveness of the fusion strategy and providing insights into the overall performance of the multimodal model.

Chapter 5 is dedicated to the development of an audio emotion recognition model. It covers the entire process, beginning with audio data processing and feature extraction. It then delves into the details of the proposed method algorithms utilized in developing the classification model. The chapter describes the training and testing procedures, experimental setup, and presents the results and corresponding discussions, offering insights into the performance and effectiveness of the model.

Chapter 6 focuses on the development of a text emotion recognition model. It outlines the steps involved in text data processing, and the algorithmic details of the classification model. Similar to Chapter 5, it presents the training, testing, and experimental aspects of the model development, followed by the results and discussions to evaluate its performance.

Chapter 7 focuses on the development of a video image emotion recognition model. It emphasizes the processing of video data and highlights the extraction of image features. The chapter provides an in-depth explanation of the algorithmic details in developing the classification model. It includes the training, testing, and experimental stages, presentation of results and discussions.

Chapter 8 serves as the conclusion and discussion section of the thesis. It summarizes the key findings of the study, including the outcomes and contributions of each chapter. Additionally, this chapter acknowledges the limitations of the research, opening avenues for future work and further exploration in the field of emotion recognition.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter reviews the theoretical background of human emotions and how they relate to the current advancement of artificial intelligence. Besides that, this chapter reviews the multimodal emotion recognition data modalities, the benchmark dataset used, strategies for handling class imbalance and the fusion strategy for integrating various data modalities. Moreover, the recent state-of-the-art discoveries in deep learning for emotion recognition are also explored in this chapter. The final section of this chapter discusses unimodal approaches to emotion recognition including speech recognition, textual and facial emotion recognition.

2.2 Emotion Recognition

Emotions are mental states arising from the changes of neurophysiological that are closely related to thought, feeling, sensation, behavioral response and level of pleasure and displeasure (Panksepp, 2004). Emotions can be expressed through interaction and socializing with other peoples and they play an essential influence in daily life. Each day, humans spend a considerable amount of time witnessing the emotions of others, understanding and interpreting the meaning of those signals as well as determining ways to respond to the complex emotional experience. Emotional awareness helps humans to build better relationships, understanding what a person and other people need and want. This is because being aware of emotions would help humans to communicate more clearly about feelings, avoid and resolve conflicts in a better way, as well as deal with past unpleasant feelings more easily.

The process of identifying human emotion is known as emotion recognition. Emotion recognition is certainly challenging due to the complex and often subtle nature of human emotions, which can vary widely between individuals and contexts. Many researchers have devoted a great amount of effort, time and energy towards understanding the existence of emotions and developing theories about how the emotions are present and occur in a human being (James, 1884). Factors such as cultural differences, personal experiences and social norms influence how emotions are expressed, making it difficult to establish universal standards for recognizing them accurately (Elfenbein & Ambady, 2002). Moreover, identifying emotions reliably also requires understanding non-verbal cues like facial expressions, tone of voice, and body language, which can be easily misinterpreted.

Nevertheless, emotion recognition plays a significant role in improving user experience, especially by enabling personalization and real-time feedback. For example, systems that understand a user's emotional state can personalize content, recommendations, and interactions. In e-commerce, if a system senses user frustration, it can offer help or suggestions to make shopping easier. Similarly, entertainment apps might recommend content that matches a user's mood, like uplifting videos when someone feels down. Real-time feedback also improves responsiveness. For example, virtual assistants and chatbots can adjust their tone or connect the user to a live agent if they sense frustration, creating a more supportive experience.

In mental health, emotion recognition offers valuable support by tracking changes in mood that could signal mental health concerns. This is especially helpful in telehealth, where artificial intelligence (AI) tools can passively monitor emotional patterns and

alert professionals if someone’s mood shifts significantly. Emotion recognition can also detect signs of anxiety or depression and suggest interventions early, helping users access support when needed.

Emotion recognition improves the adaptability, engagement and safety of AI interactions. The systems make users feel more connected by tailoring responses to their emotions, fostering trust and engagement. Moreover, in education for example, emotion recognition can detect confusion or frustration in students. Emotion recognition also improves safety in autonomous systems such as in-car AI, by monitoring drivers' stress or fatigue levels and recommending breaks or adjustments to ensure everyone's safety.

2.2.1 Types of Emotion in Recognition Systems

In emotion recognition systems, emotions are categorized in order to identify and classify emotional states accurately. Common categories include basic emotions, complex emotiAAons, and dimensional models. Each type serves different recognition needs and application goals. Previous studies on these categories highlight varying approaches and their implications for developing effective recognition systems.

Basic emotions are universal expressions that serve as the foundation for many emotion recognition systems. Paul Ekman’s model which includes happiness, sadness, anger, fear, disgust, and surprise has been widely used besides proposing that these basic emotions are universally recognized across cultures (Ekman, 1992). This concept has been fundamental since research has shown that these emotions exhibit

recognizable, consistent patterns in tone, facial expressions, and physiological reactions (Kohler et al., 2004). These simple classifications allow for reliable detection in applications like customer service, entertainment, and virtual assistants. Recent research continues to leverage these basic emotions due to their consistency especially in studies on human-computer interaction where basic emotional cues are critical for creating responsive interfaces(Moin et al., 2023) .

In addition to basic emotions, complex emotions such as jealousy, guilt, and pride present a unique challenge for recognition systems as they are context-dependent and often involve a combination of basic emotions. For instance, envy can be a combination of anger and sadness, while pride merges happiness with a feeling of accomplishment. Early research by (Shaver, 1987) identified these emotions as higher-order categories that involve personal experiences and social context that make them less universally recognizable than basic emotions. As a result, recent studies explore context-aware systems that use advanced natural language processing and multimodal inputs to interpret complex emotions more effectively. (J. Li et al., 2022). These approaches are necessary as complex emotions often lack distinct facial or vocal cues and rely on contextual subtleties, requiring more nuanced processing techniques.

Apart from that, dimensional models present another approach that focus on representing emotions as points along continuous scales rather than discrete categories. Russell's Circumplex Model organizes emotions within a circular structure using two main dimensions which are arousal (energy level) and valence (positivity or negativity)(Russell, 1980). This framework allows recognition systems to consider the intensity and nature of emotional states such as differentiating between calmness

and excitement which share positive valence but differ in arousal levels. The Valence-Arousal-Dominance (VAD) model extends this by adding a third-dimension dominance that describe the level of control within an emotional state (Mehrabian, 1996). Dimensional models have gained traction in recent studies due to their ability to capture subtle emotional differences and increasingly applied in systems using neural networks to detect emotional gradients in voice, facial expressions and physiological response. These models are particularly effective for real-time monitoring systems where nuanced emotional responses are essential (Nandi & Xhafa, 2022) .

2.2.2 Emotion Artificial Intelligence (AI)

The ability to recognize and regulate one's own emotions, as well as the emotions of others is referred to as emotional intelligence. Emotional intelligence is present almost naturally to human and people use instinct to elicit reactions. Humans can easily perceive the emotions and moods of people around them and this kind of intelligence guides humans on how to act with the right behavior in certain situations. However, for a computer or machine, interpreting emotions and responding to them is relatively a difficult task. It then leads to a question on how can a machine communicate effectively with a human if it does not know the emotional state and feelings of a human being?

Recent advancements in AI particularly in deep learning and neural networks, have greatly improved the accuracy of emotion recognition systems by enabling them to detect and analyze complex, subtle patterns in data. Traditional machine learning approaches focused primarily on explicit cues like facial expressions or vocal tone,

which limited their accuracy in diverse, real-world contexts (Cambria et al., 2017). Deep learning models, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have expanded the capability of these systems by processing large volumes of nuanced data across modalities such as audio, text, and video. CNNs are particularly effective at identifying spatial patterns in images and videos. Meanwhile RNNs are skilled at capturing sequential dependencies in speech and text data, both critical in accurately interpreting emotional cues (Tzirakis et al., 2017).

AI-powered emotion recognition systems have evolved to integrate these advancements, resulting in sophisticated multimodal frameworks that analyze diverse data streams in parallel. For instance, a multimodal system might process audio, text, and video inputs simultaneously, synthesizing information from voice tone, word choice and facial expressions thus, derive a more holistic emotional understanding (Poria et al., 2017). As a result, recent studies demonstrate substantial improvements in emotion recognition accuracy in real-world applications, from virtual assistants to mental health monitoring where understanding subtle emotional cues is essential (A. Gupta & Batra, 2023) . By leveraging deep learning techniques, AI has enabled emotion recognition systems to offer unprecedented insights, enhancing their reliability and applicability across a wide range of contexts.

2.3 Multimodal Emotion Recognition

Emotion recognition research has largely focused on single-modal recognition such as facial expression recognition, speech recognition, limb physiological signal recognition and text emotion recognition. However, it may result in a lower accuracy of emotion recognition due to lack of single-modal emotional information and

vulnerability to various outside influences. For example, speech is susceptible to interference from background noise and facial expressions are easily obstructed. As a result, multimodal systems are utilized to compensate for the low accuracy rates of certain emotions (Malygina et al., 2019).

In 1997, multimodal emotion data fusion and recognition was firstly introduced by Bigun and Duc (Duc et al., 1997). They combined voice as well as facial data and proposed a statistical strategy based on the Bayesian theory. These days, combining various inputs, such as image, video, and speech, has emerged as a revolutionary method to identify a person's precise emotion. It can be observed when significant advancements have been made in the study of multimodal emotion data fusion and recognition (Jiang et al., 2020).

2.3.1 Multimodal Fusion Techniques

The difficulty in fusing multimodal emotion data is mainly caused by emotion cognition and analysis after combining diverse emotion data modalities from different sources and distinct time scales (Poria, Cambria, et al., 2016). Applying the right fusion techniques to get the best accuracy depends in part on the particular case (Datcu & Rothkrantz, 2015). Multimodal fusion techniques can be broadly categorised into two categories which are feature-level fusion and decision-level fusion. The detailed explanation of each multimodal fusion technique is presented below.

A. Feature-level Fusion (Early Fusion)

The conventional way of integrating multiple data before completing the analysis is called feature-level fusion also known as early fusion. With the use of this technique,

features from each modality are combined to create a new feature vector. It can be difficult to synchronize data sources when one is discrete and the others are continuous. Therefore, integrating data sources into a single feature vector is one of the biggest challenges in feature-level early data fusion. In addition, the new feature vector has a tendency to have a higher dimension. To solve this problem, the dimension reduction method will be used before a classifier is finally used to detect the emotion.

Feature-level early data fusion is based on the notion that various data sources are conditionally independent of one another. Nevertheless, (Sebe, 2005) argues that this notion is not always correct because different modalities can have features that are highly correlated, for instance video and depth cues. This allows for the fusion of feature layers using the mutual relations between the different modalities. The use of feature-level early data fusion however has some drawbacks and one of them is that a significant amount of information will be taken out from the modalities to create a common ground prior to fusion. Apart from that, it is difficult to achieve time synchronism between various modalities. As the number of modalities increases, learning the relevance across modalities will even get harder (Poria et al., 2017).

Despite the shortcomings in feature-level early data fusion, there have been many previous studies that have benefited from the fusion approach. For example, a study from (Chao et al., 2015) employed LSTM to train the fused features in the feature layer prior to the implementation of linear regression for emotion recognition. Moreover, study from (Tzirakis et al., 2018) fused the 1280 dimensions of phonetic features with 2048 dimensions of image features, and obtained 3328 dimensions of

feature vector. Meanwhile in (Nguyen et al., 2018), a new approach based on the integration of visual and phonetic data according to the bilinear pooling theory was proposed. The DBN was then used to train the fused features, and a softmax classifier was then used to recognize emotions. A novel feature fusion technique based on the Relational Tensor Network that fused text, audio, and image was proposed in (Sahay et al., 2018) and the CMU-MOSEI dataset was used to validate the effectiveness of the method proposed. Study from (Hazarika, Gorantla, et al., 2018) fuse audio and text features to create new emotional features by using a self-attention method. Apart from that, both studies from (S. Zhang et al., 2017)(Ma et al., 2019) employed the DBN network to fuse the emotional features of speech and facial expression, then trained the network to learn new emotional features after the two modalities have been fused.

B. Decision-Level Fusion (Late Fusion)

The popularity of ensemble classifiers, in which data sources are employed separately and combined at the decision-making stage, served as an inspiration for decision-level fusion also known as late fusion (Kuncheva, 2014). In general, the decision-level late fusion process will involve the implementation of a respective classifier that takes into account multimodal features such as text, image and audio features for emotion recognition. The results of single-modality emotion recognition are then combined using an algebraic combination rule to optimize the final result of recognition accuracy. In particular, when the data sources contain significantly different sampling rates, data dimensionality, and unit of measurement from one another, this strategy is easier to use than the features-level early fusion method. Despite the fact that a study by (Ramachandram & Taylor, 2017) argues that there is insufficient proof to say that decision-level late fusion outperforms feature-level early fusion, decision-level late

fusion frequently produces favourable performance since errors from various models are handled independently, making errors uncorrelated.

There are a number of rules for determining the best way to combine all of the individually trained models. Some of the popular decision-level late fusion rules include bayes rule, max-fusion, average-fusion and weighted average fusion. In a study from (B. Sun et al., 2015), the weighted product rule was implemented to fuse the results of audio and image recognition at a decision layer. The weights in the fusion network are multiplied by the probabilistic values for each class of each feature that were previously obtained before the values belonging to the same class are added together. Finally, the final label will be selected based on the highest probabilistic value. Apart from that, a study from (Y. Huang et al., 2017b) implemented the sum rule and production rule to fuse the recognition from EEG and facial expression classification. Meanwhile, in a study from (Fan et al., 2016) used weighted summing as part of the decision-level fusion technique to combine the predicted scores obtained from the various network models.

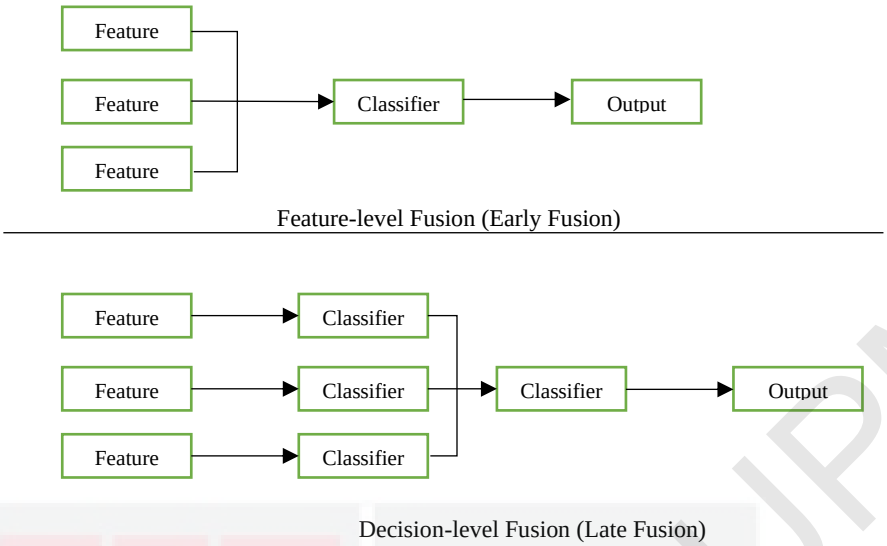

Figure 2.1 : llustration of Feature-Level Fusion vs Decision-Level Fusion

Figure 2-1 presents the illustration of feature-level fusion and decision-level fusion. In feature-level fusion, features from different modalities (e.g., audio, text, and video) are combined into a single feature vector, which is then processed by a single classifier. This allows the model to learn shared representations across modalities but may struggle when modalities differ significantly in scale or format. In contrast, decision-level fusion processes each modality independently, with separate classifiers for each; the resulting predictions are then combined using a final classifier or fusion rule to make the overall decision. This approach is advantageous when modalities have unique characteristics or when handling missing data, as predictions from available classifiers can still contribute to the final output.

2.3.2 Recent Advances in Multimodal Emotion Recognition

Recent advances in multimodal emotion recognition have significantly benefited from deep learning techniques, particularly through CNN and RNN architectures. These models have improved the ability to capture complex patterns within and across

modalities such as text, audio, and video, making it possible to develop systems that interpret human emotions more accurately. CNNs, for example, are widely used for processing visual data, making them effective for emotion recognition in video-based modalities. By applying convolutional layers to extract spatial features from facial expressions, CNNs capture fine-grained details crucial for identifying subtle emotional cues (Zhou et al., 2020) .However, CNNs are generally limited in capturing temporal dependencies independently since they primarily analyze spatial features on a frame-by-frame basis. To address this, CNNs are frequently combined with RNNs, which are better suited for processing sequential data by allowing multimodal systems to integrate both spatial and temporal information (Fan et al., 2016).

RNNs, particularly Long Short-Term Memory (LSTM) networks, have shown effectiveness in handling sequential data in audio and text modalities. These networks excel at capturing temporal dependencies, enabling emotion recognition systems to understand how emotions evolve over time (Tzirakis et al., 2018). However, RNNs face challenges in managing long-range dependencies especially in real-time applications. This has led to the adoption of Transformer models that employ self-attention mechanisms to capture dependencies across both time and modalities without the sequential constraints of RNNs (J. Li et al., 2020).

Advances in architectures like transformers and attention mechanisms have further pushed the boundaries of multimodal emotion recognition. Transformers, with their self-attention mechanism, have revolutionized text and audio processing by allowing models to focus on key parts of input sequences, capturing intricate contextual relationships (Vaswani et al., 2017). In emotion recognition, attention mechanisms

enable systems to weigh the importance of different cues within a dataset, such as distinguishing between emotional inflections in speech or subtle facial expressions in video. Research has shown that models using attention and transformers significantly outperform earlier models in tasks requiring fine-grained emotional analysis (J. Zhang et al., 2022). These methods provide flexibility in recognizing blended emotions or subtle shifts, such as sarcasm in text or fatigue in facial expressions, which are often difficult for traditional models to interpret accurately.

To enhance performance further, attention-based mechanisms have been integrated with CNNs and RNNs, allowing models to focus on the most informative features in each modality. Cross-modal attention, for instance, helps these models dynamically prioritize cues from one modality over another based on the emotional context (X. Xu et al., 2020). However, attention mechanisms can still suffer from imbalances, where dominant modalities overshadow others, reducing the robustness of the system. To avoid this, multi-head attention has been explored to provide more balanced representation across multiple dimensions, though this adds complexity and computational demand (J. Zhang et al., 2022).

Several multimodal emotion recognition systems have been proposed using various innovative approaches. For instance, (Majumder et al., 2018) presented a Gated Recurrent Unit (GRU) based model to capture contextual information besides utilizing CNN to extract features from text and visual modalities. (Poria et al., 2018) presented an LSTM-based model to capture contextual information from videos and a bidirectional LSTM+SVM-based system that applies CNNs for extracting features from text and visual modalities. Another system employed a hierarchical fusion

strategy, achieving high accuracy by combining text, audio, and video modalities (Gu et al., 2018). Similarly, (Tripathi et al., 2019) used neural networks with final-layer fusion for more robust results, while (Sreevidya et al., 2024) explored mid-level fusion to integrate acoustic and textual features

2.3.3 Multimodal Data

Multimodal emotion recognition involves the integration of multiple modalities such as audio, text, and visual cues to accurately detect and interpret human emotions. Audio data captures features like tone, pitch, and intonation, which can provide insights into a person's emotional state through speech patterns. Textual data from sources such as chat logs or transcripts provides an additional layer by extracting emotional content from written communication using sentiment analysis and natural language processing techniques. Meanwhile, visual modalities, including facial expressions, body language, and gestures, offer rich cues for emotion recognition, as they can convey a wide range of feelings from happiness and surprise to anger and fear.

In multimodal emotion recognition, datasets play an important role and serve as the repository of information that are required for emotion recognition. The majority of procedures used to gather multimodal data for emotion recognition include inducing certain emotions in test subjects through the use of videos or other mediums. The corresponding data is labelled and recorded when a desired emotion is produced. The obtained data is archived in a dataset known as the induced or acted dataset. Apart from this, there are certain works that record the unprompted emotions of the test

subjects and the dataset is known as spontaneous dataset. Table presents several benchmark datasets frequently used in multimodal emotion recognition:

Table 2.1 : Benchmark Datasets in Multimodal Emotion Recognition

Dataset	Description	Modality Types	Key Features
IEMOCAP (Busso et al., 2008)	Interactive Emotional Dyadic Motion Capture, designed for emotion recognition in natural conversations.	Audio, Video, Text	- Consists of 10 actors (5 males, 5 female) - Contains 12 hours of annotated data - Supports 5 primary emotions (happy, sad, angry, neutral, and frustrated)
AFEW (Dhall et al., 2012)	Affective Expression in the Wild, focused on emotion recognition from spontaneous facial expressions in videos.	Audio, Video	- Derived from movies, providing realistic scenarios - Annotated for 7 emotions (happy, sad, surprise, disgust, anger, fear, neutral) - Challenges include varying lighting and expressions
BAUM-1 (Zhalehpour et al., 2016)	A dataset for multimodal emotion recognition from audiovisual recordings in natural settings.	Audio, Video	- Consists of diverse emotional interactions - Focuses on real-life conversations - Provides annotations for 5 basic emotions
CMU-MOSEI (Zadeh et al., 2018)	The CMU Multimodal Opinion Sentiment and Emotion Intensity dataset, aimed at understanding sentiment and emotion in video clips.	Audio, Video, Text	- Contains over 3000 annotated videos - Focuses on sentiment analysis alongside emotion recognition - Rich annotations for emotions and sentiment intensity

2.3.4 Class Imbalanced in Multimodal Data

Acquiring a large, well-annotated multimodal dataset remains a major challenge in developing effective multimodal emotion recognition systems. Limited availability of

audio, text, and visual data makes training deep neural networks particularly difficult, as these models require substantial amounts of training data for effective learning and evaluation (Tsalera et al., 2021). Moreover, insufficient data often results in class imbalance, where certain categories are underrepresented, leading to a skewed distribution. This imbalance hinders the model's ability to learn meaningful patterns from minority classes, ultimately reducing classification accuracy and increasing bias in predictions (Abayomi-Alli et al., 2022).

To address the issue of class imbalance, researchers commonly adopt three main strategies: the data-level approach (Guzmán-Ponce et al., 2021), the cost-sensitive approach (Devi et al., 2022), and the algorithm-level approach (Ganaie et al., 2021). Among these, the data-level approach is the most widely used, as it involves preprocessing techniques to balance the dataset before classification. This includes oversampling methods, which increase the number of minority class samples (Koziarski et al., 2021) and undersampling methods, which remove a portion of the majority class data (Chennuru & Timmappareddy, 2022). A key advantage of the data-level approach is its versatility, as it can be applied to any classifier (Elreedy et al., 2023).

Various strategies have been proposed to improve the handling of class-imbalanced data. For instance, a Naïve Bayes classifier has been widely used in undersampling techniques (Aridas et al., 2019). (H.-S. Choi et al., 2021) introduced an innovative three-component framework consisting of a discriminator, generator, and classifier, incorporating decision boundary regularization. A key aspect of this approach is the joint training of a generator alongside a classifier. (Xie et al., 2021) proposed a novel

undersampling technique that utilizes consecutive density peaks to iteratively remove majority class samples, thereby refining the dataset. study (Zheng et al., 2021) explored genetic algorithm-based methods to automatically determine sample ratios for oversampling, undersampling, and hybrid techniques. Additionally, (Mayabadi & Saadatfar, 2022) introduced two density-based sampling techniques—density-based undersampling (DB_US) and density-based hybrid sampling (DB_HS)—designed to eliminate class overlap and create a more balanced and normalized data distribution.

Recently, the Synthetic Minority Over-sampling Technique (SMOTE) has garnered significant attention from researchers. SMOTE was created by (Chawla et al., 2002a) and is one of the most widely used over-sampling techniques. The SMOTE method generates synthetic data by applying linear interpolation between minority class samples. As a widely used oversampling technique, SMOTE has been extensively applied in various studies. (Dablain et al., 2022) introduced a novel oversampling algorithm based on SMOTE, specifically designed for deep learning models.

(Elyan et al., 2021) proposed a hybrid approach called CDSMOTE, which combines class decomposition with oversampling to mitigate the dominance of majority class samples. Unlike general undersampling methods, this approach preserves majority class samples while achieving a more balanced dataset.

(Thejas et al., 2022) advanced SMOTE by integrating it with a Kalman filter, effectively removing noisy samples from datasets that include both original and synthetic data. To further address noise-related issues, they developed IR-SMOTE, an oversampling algorithm that first organizes majority class samples before applying k-

means clustering. The kernel density estimation method is then used to distribute synthetic samples across clusters appropriately.

Nevertheless, one notable limitation of SMOTE-based methods is their dependence on the choice of classifier. Since SMOTE creates synthetic samples to address class imbalance, its success relies on how well the classifier can handle these generated instances. Certain classifiers may have difficulty integrating the synthetic data effectively, which could impact overall classification performance

Table 2.2 : Summary of Multimodal Emotion Recognition Fusion Approach

Author	Dataset	Fusion	Mode	Modality & Features	Method
(Fan et al., 2016)	AFEW	Decision-level (Late fusion)	Bimodal	- Audio: SVM - Video: CNN-LSTM, 3DCNN	Weighted Sum for Decision Level
(S. Zhang et al., 2017)	RML eNTERFACE BAUM-1	Feature-Level (Early Fusion)	Bimodal	- Audio: CNN - Visual: 3DCNN	Deep Belief Network (DBN) SVM
(Nguyen et al., 2018)	eNTERFACE FABO	Feature-Level (Early Fusion)	Bimodal	- Audio: Cascaded 3DCNN + Deep Belief Network (DBN) - Video: Cascaded 3DCNN	Multimodal compact bilinear pooling (MCB) Deep Belief Network (DBN)
(Sahay et al., 2018)	CMU-MOSEI	Feature-Level (Early Fusion)	Bimodal	- Audio: MFCC, i-Vector Modeling Text: Lexicons, lexico-syntactic fine- - grained, contextualized embeddings	Relational Tensor Fusion Network LSTM
(Hazarika, Gorantla, et al., 2018)	IEMOCAP	Feature-Level (Early Fusion)	Bimodal	- Audio: OpenSMILE, MFCC - Text: CNN	Self- Attention
(Hazarika, Poria, et al., 2018)	IEMOCAP	Feature-Level (Early Fusion)	Trimodal	- Audio: openSMILE Visual: 3DCNN - Text: CNN	Simple concatenation GRU
(Poria et al., 2018)	MOUD IEMOCAP	Feature-Level (Early Fusion)	Trimodal	- Audio: openSMILE - Visual: 3DCNN - Text: CNN	Bc-LSTM SVM
(Majumder et al., 2018)	CMU-MOSI IEMOCAP	Feature-Level (Early Fusion)	Trimodal	- Audio:openSMILE - Video:3DCNN - Text: CNN, Word2Vec	GRU

Table 2.2 : Continued

Author	Dataset	Fusion	Mode	Modality & Features	Method
(Ma et al., 2019)	RML eNTERFACE BAUM-1	Feature-Level (Early Fusion)	Bimodal	- Audio:2DCNN - Visual: 3DCNN	Deep Weighted Fusion Deep Belief Network (DBN)
Tripathi et al., n.d., 2019)	IEMOCAP	Decision-level (Late fusion)	Trimodal	- Audio: BiLSTM - Visual: 2DCNN - Text: LSTM	Concatenate fully connected layers
(Singh & Fang, 2020)	IEMOCAP	Feature-Level (Early Fusion)	Bimodal	- Audio: CNN RNN - Video:3DCNN	Concatenate fully connected layers
(Tzirakis et al., 2021)	SEWA	Feature-Level (Early Fusion)	Trimodal	- Audio: High Resolution Network - Video: High Resolution Network - Text: Transformer	Attention LSTM
(Santoso et al., 2022)	IEMOCAP	Feature-Level (Early Fusion)	Bimodal	- Audio: Spectogram, BLSTM, Self-Attention - Text: BLSTM, Self-Attention	Self-Attention Weight Correction
(G. Zhao et al., 2023)	IEMOCAP MELD	Feature-Level (Early Fusion)	Bimodal	- Audio: CNN, MFCC, Auxiliary Acoustic - -Text:BERT	CNN

Table 2-2 summarized the fusion approach implemented in the previous studies for multimodal emotion recognition. Many previous researchers showed a preference for implementing feature-level early fusion over decision-level late fusion in multimodal emotion recognition. While feature-level early fusion may be easier to train when compared to late fusion, it frequently encounters overfitting issues due to its large number of parameters. This problem becomes even more pronounced when dealing with trimodal systems that combine features from three different modalities (Hazarika, Poria, et al., 2018; Majumder et al., 2018; Poria et al., 2018; Tzirakis et al., 2021) . As a result, emotion labels tend to get misclassified and harmonizing features with varying dimensions becomes a formidable task. Although it's important to acknowledge that decision-level late fusion has its own shortcomings, it is worthwhile to explore this fusion strategy further as a means to address the weaknesses associated with early fusion.

2.4 Unimodal Approaches in Emotion Recognition

The unimodal emotion recognition approach relies on data from a single source or modality, such as voice tone from speech, text or facial expression to infer emotional states. This approach typically involves analyzing features specific to the chosen modality to train machine learning models for emotion classification. While unimodal methods offer simplicity and ease of implementation, the quality of performance remains crucial especially in scenarios where late fusion is employed to achieve optimal and accurate final results.

2.4.1 Speech Emotion Recognition (SER)

Verbal communication is the most natural and effective means of human interaction. This has led researchers to consider speech signals as a rapid and efficient medium for human-computer interaction. Consequently, Speech Emotion Recognition (SER) has become a crucial aspect of Human-Computer Interaction (HCI), enabling computers to recognize and respond to human emotions. Over the past two decades, SER has been widely applied in various domains requiring human-computer interaction, including call center conversations, online tutoring, medical analysis, and more (El Ayadi et al., 2011; Schuller, 2018).

The primary objective of SER is to automatically identify a speaker's emotional state based on vocal cues. In essence, SER encompasses a set of techniques designed to analyze and classify speech signals to uncover the emotions embedded within them. Therefore, a high-quality of annotated emotional speech database is necessary for SER to work successfully (Ververidis & Kotropoulos, 2006). The labelled data need to be pre-processed before the extraction of relevant features. Early approaches emphasized manually extracting features and using standard methods like Gaussian Mixture Models (GMM), Dynamic Time Warping (DTW), and Hidden Markov Models (HMM) (Karpagavalli & Chandra, 2016). In the deep learning era, methods like CNN, RNN have been widely implemented and achieved great performance. After the pre-processing and feature extraction are done, all of the features are then forwarded to a suitable classifier for classification.

A. SER Features

Features are the most important element in classification. It is necessary to distill the original data to the characteristics that are truly crucial and significant for classification. Researchers have investigated and incorporated a variety of features for SER. As for now, there is no standard method for extracting speech features and specific classifiers yet. Speech is a continuous signal with varying lengths that transmits both information and emotion. Therefore, it really depends on one's objective and target in order to determine which local or global features are best extracted from the speech signal. The temporal dynamics represent the local features which are also known as short-term features of segmental features. Meanwhile, comprehensive statistics like minimum and maximum values, means, and standard deviation are used to represent global features which is also known as long term or supra-segmental features.

The local and global features for SER are analyzed under four following categories: Prosodic feature, Teager Energy Operator (TEO), Voice Quality features and Spectral features (Wani et al., 2021).

i. Prosodic Feature

- Based on previous studies, there is a correlation between prosodic features and human emotional states (Lee & Narayanan, 2005; Nwe et al., 2003b; Schroder & Cowie, 2006). Prosodic features include intonation and rhythm which can be perceived by humans from words, sentences, syllables and expression (Swain et al., 2018). According to (Koolagudi & Rao, 2012), prosodic features are mainly categorized in three categories which are pitch, intensity and energy. (Wani et al., 2021) mentioned in their study that the most frequent use of prosodic features is based on fundamental frequency, energy, and duration. Statistical value of the prosodic features such as mean, maximum and

minimum values and range contain various emotional information and are very useful in recognizing emotions (Wu & Liang, 2010).

ii. Teager Energy Operator (TEO)

- Teager Energy Operator (TEO) was first proposed to detect stress in speech by Teager and Kaiser (Kaiser, 1990; Teager, 1980). According to Teager, a non-linear vortex-airflow interaction occurs in the human vocal system to produce speech. In a stressful situation, the speaker's muscles tense up and change the airflow during sound production.

iii. Voice Quality

- The individual voice characteristic has been defined as the voice quality features. Due to the significant correlation between voice quality and the emotional content of the speech, voice quality features become the centre of many speeches processing applications, including speaker identification and emotion recognition (Cowie et al., 2001). Vocal quality features include format frequency, bandwidth, glottal parameter, harmonic to noise ratio, jitter, and shimmer.

iv. Spectral features

- The sound produced when a person speaks is filtered by the shape of the vocal tract. The frequency domain provides a good representation of vocal tract characteristics (Koolagudi & Rao, 2012). The Fourier Transform is used to transform a time domain signal into a frequency domain signal in order to extract spectral features. There are many different spectral features that have been described in previous studies including Mel Frequency Cepstral Coefficient (MFCC) (Mansour & Lachiri, 2017), Log-Frequency Power Coefficients (LFPC) (Nwe et al., 2003a), Gammatone Frequency Cepstral Coefficients (GFCC) (Chatterjee et al., 2015) and etc.

- MFCC is the most extensively employed spectral feature especially in this era of deep learning (Akçay & Oğuz, 2020). The mel scale is a logarithmic scale built on the concept that identical lengths on the scale correspond to the same perceptual distance. It appears that humans experience sound logarithmically.

Lower frequencies are easier to distinguish from higher frequencies. For instance, despite the fact that the distance between the two pairs is the same, it will be difficult for a human to distinguish between frequencies of 500 and 1000 Hz and 10000 and 10500 Hz, respectively. As a result, a mel-spectrogram is a spectrogram in which the frequencies are translated to the mel scale. The following formula converts frequency (f) to mel scale (m):

$$m = 2595 . log\ (1 + \frac{f}{500})$$

(Equation 5)

The information about the rate of change in a signal's spectral bands is provided by ceptrum. The ceptrum is produced by performing an inverse Fourier transform (IFT) on the estimated signal spectrum's logarithm and MFCC is basically the set of coefficients that make up the mel-frequency cepstrum.

- In order to optimize speech recognition, it is necessary to comprehend the power spectrum's dynamics, such as the trajectories of MFCCs over time. Therefore, it is encouraged to include delta (differential) and delta-delta (acceleration) features of MFCC features as well. The following formula is used to calculate the delta coefficients:

$$dt = \frac{\sum Nn = 1n(ct + n - ct - n)}{2\sum Nn = 1n2}$$

(Equation 6)

Similar computations are done for the delta-delta coefficients, but they use the differential rather than the static coefficients.

B. Deep Learning Approach in SER

Researchers have explored various classifiers for Speech Emotion Recognition (SER) to enhance accuracy and performance. Traditional machine learning approaches relied on classifiers such as Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), Artificial Neural Networks (ANN), and K-Nearest Neighbors (K-NN), which were applied after extracting speech features (El Ayadi et al., 2011). Since 2013, DL methods have gained increasing popularity in emotion recognition (Y. B. Singh & Goel, 2022). Today, most SER models utilize DL techniques, achieving improved accuracy and efficiency.

Over the years, CNN and LSTM have emerged as the most commonly used architectures. In 2014, researchers (Z. Huang et al., 2014; Mao et al., 2014) employed CNN with spectrograms for emotion recognition. The following year, S. Chen & Jin, 2015) implemented an RNN classifier, attaining an 81% accuracy rate on the RECOLA database. In 2017, (W. Zhang et al., 2017) applied a Deep Belief Network (DBN) classifier to the CASEC dataset, achieving 94.60% accuracy. (J. Zhao et al., 2018) later obtained 92.71% accuracy using a Deep Convolutional Neural Network (DCNN) on the EMODB dataset.

By 2020, CNN models continued to yield strong results, with (Anvarjon et al., 2020) achieving 95% accuracy on the EMODB dataset. The trend toward hybrid approaches has also increased over time. For instance, (J. Zhao et al., 2019) integrated 2D CNN with LSTM, extracting features from spectrograms and evaluating performance on the EMODB and IEMOCAP datasets, achieving 95.89% and 89.16% accuracy, respectively.

Table 2-3 summarized the SER approach implemented in the previous studies. Based on the table presented, a noticeable trend among researchers since 2019 has been the incorporation of attention mechanisms and BILSTM approaches into their work. While these techniques have yielded accuracy levels similar to other established methods in deep learning, they bring several additional advantages particularly in terms of efficiency in capturing long-range dependencies within temporal sequences of speech features and focusing on relevant input components. Notably, the IEMOCAP database has garnered significant attention from researchers in the field of SER (Santoso et al., 2022; Tripathi et al., 2019; Xu et al., 2019; J. Zhang et al., 2022; J. Zhao et al., 2018, 2019). Many researchers focused towards utilizing spectral features, rather than other categories speech features during this deep learning era of SER.

Table 2.3 : Summary of SER in Previous Studies

Author	Dataset	Feature	Approach						Imbalanced Class Data Handler	Remark
			CNN	LSTM	ATT	BILSTM	DBN	SVM		
(Z. Huang et al., 2014; Mao et al., 2014)	SAVEE Emo-DB DES MES	Spectral (Specto-gram)	/					/	-	Able to extract affect-salient features by disentangling emotions from other factors such as speakers and noise.
(S. Chen & Jin, 2015)	RECOLA	Spectral		/					-	Single direction of LSTM is good enough for dimensional emotion regression compared to BiLSTM as it is more time consuming and prone to overfitting.
(Satt et al., 2017)	IEMOCAP	Spectral	/	/					Two-step prediction	Enables effective handling of background non-speech signals such as music and crowd noise.
(W. Zhang et al., 2017)	CASEC	Spectral (MFCC) Prosodic (Pitch)					/		-	Combined several kinds of speech features to improve the performance of DBN
(Zhao et al., 2018	IEMOCAP Emo-DB	Spectral (log mel spectro-gram)	/						-	CNN architectures can learn hierarchical features and model high-level abstractions of the emotional information from the raw audio clips and log-mel spectrograms
(J. Zhao et al., 2019)	IEMOCAP	Spectral	/	/					-	CNN LSTM networks able to learn local information and long-term contextual dependencies from the inputted features

Table 2.3 : Continued

Author	Dataset	Feature	Approach						Imbalanced Class Data Handler	Remark
			CNN	LSTM	ATT	BILSTM	DBN	SVM		
(Xu et al., 2019)	IEMOCAP	Spectral (MFCC)				/			-	Attention mechanism able to learn alignment in Automatic Speech Recognition
(Tripathi et al., 2019)	IEMOCAP	Spectral (Mel Spectro-gram)		/	/	/		/	-	AttBiLSTM allowed minimal pre-processing and no complex loss functions or noise injection into the training
(Santoso et al., 2022)	IEMOCAP	Spectral (MFCC)			/	/			-	Able to adjust the importance weights of speech segments and words with a high possibility of ASR error
(J. Zhang et al., 2022)	IEMOCAP	Spectral		/	/				Up-sampling	MFCC has been shown to be more successful because it is a highly compressible representation of the speech signal, making it more "biometric."
(L.-W. Chen & Rudnicky, 2023)	IEMOCAP	Prosodic	/						-	Fine-tuning strategies for the Wav2vec 2.0 model, including Vanilla Fine-Tuning (V-FT), Task Adaptive Pretraining (TAPT) and a novel method called P-TAPT, to enhance performance in speech emotion recognition tasks

However, instead of relying solely on spectral features like mel spectogram and MFCC, it is important to include features related to temporal context as emotions in speech often exhibit temporal dynamics. Moreover, most of the prior studies inadequately addressed the handling of imbalanced class and data distributions, which could result in problems like overfitting and a high rate of misclassified test cases for specific emotion class labels (J. Zhang et al., 2022). While study from (Satt et al., 2017) introduced a two-step prediction approach as a solution, it comes with certain drawbacks, such as the need for additional hyperparameter tuning to ensure the effectiveness of both steps, which can be a challenging and time-consuming task.

2.4.2 Textual Emotion Recognition (TER)

Textual Emotion Analysis (TER) is a process of extracting and interpreting human emotional states in texts. Compared to sentiment analysis which primarily evaluates subjective views from the viewpoint of sentiment polarity, TER involves the identification of more specific emotional states that refers to a wide range of mental states such as sadness, happiness, anger and fear (Deng & Ren, 2021). TER plays important roles in many NLP applications such as e-commerce, online teaching, mental-health management, public opinion analysis and many more.

In contrast to speech recognition and facial expression analysis, a text sentence lacks the ability to define itself properly and it can be very challenging to determine the text's emotions due to their complexity and ambiguity. The difficulty arises from the fact that each word can have a varied meaning and morphological form, making it challenging to identify the emotion in a particular text (Bharti & Babu, 2018). Over the years, researchers proposed various techniques of recognizing emotions from text

including keyword-based, lexical affinity-based, learning based and hybrid models (Chopade, 2015). Each of the systems however, had a lot of drawbacks such as inadequate lists of all emotions, insufficient word vocabularies in lexicon, poor ability to extract context information, loose semantic feature extraction and a high rate of misclassifications.

For the past decade since 2012, deep learning-based techniques (DL) have been widely and frequently applied in TER. DL methods focus on feature learning as their main objective. It is essentially a way of learning complex feature representation based on original feature input using multi-layer nonlinear processing. Word Embedding is a technique that captures the inter-word semantics by representing individual words as real-valued vectors in a lower-dimensional space. The main role of this Word Embedding technique is to extract features from text and feed them as input into a DL model. The pre-trained word embedding has become a solution to the sparse feature and high-dimensional representation issues faced by standard bag-of-words models. Some well-known word embedding models like Word2Vec (Mikolov et al., 2013), Glove (Pennington et al., 2014), CoVe (McCann et al., 2017), ELMo (Peters et al., 1802) and BERT (Devlin et al., 2018) are extensively used and have been proved to improve the performance of NLP.

DL neural network architecture and knowledge-enhanced representation emphasize on developing and training an efficient network to automatically learn multi-layered emotional features. A variety of deep learning (DL) methods, including recurrent neural networks (RNN), long-short term memories (LSTM), gated recurrent neural networks (GRNN) and other variations have been developed in recent years to extract

sequence information and relevant semantic information from text features. For example, study from (Ghosal et al., 2019) introduced the Dialogue Graph Convolutional Network (DialogueGCN) for conversational emotion recognition which consist of three essential components; sequential context encoder, speaker-level context encoder and emotion classifier. Study from (D. Zhang et al., 2019) presented Graph-based Convolutional neural Network towards Conversations also known as ConGCN to cater both context-level and speaker-level dependence for emotion recognition. In order to identify emotion in a conversation, study from (Majumder et al., 2019) presented the DialogueRNN neural architecture, which is built based on RNN and also implemented CNN to extract the textual feature of each speech.

Apart from that, study from (Jiao et al., 2019) proposed Hierarchical Gated Recurrent Unit (HiGRU) framework with two Bi-GRU. The upper-level Bi-GRU was used to learn the contextual utterance embedding while the lower-level Bi-GRU was employed to learn the individual utterance embedding. Study from (J. Li et al., 2020) introduced the transformer-based HiTrans model, which consists of two hierarchical transformers. The HiTrans framework was initially proposed in (Q. Li et al., 2020) to handle utterance-level emotion recognition in dialogue systems. The framework employed both lower and upper-level transformers to model word-level input and capture the contexts of utterance-level embedding besides, implement BERT to improve individual utterance embeddings.

Table 2.4 : Summary of TER in Previous Studies

Author	Dataset	Input Representation				Approach									Remark
		GloVe	Word2Vec	BERT	OTHERS	CNN	GCN	GRU	BIGRU	LSTM	BILSTM	TRANSFORMER	MLP	ATTENTION	
(Ghosal et al., 2019)	IEMOCAP AVEC MELD	/				/	/								Encoder module in the form of a graphical network to capture speaker dependent contextual information
(D. Zhang et al., 2019)	MELD	/				/					/				Tackle context sensitive dependence in multi speaker conversation
(Majumder et al., 2019)	IEMOCAP AVEC	/						/							Recurrent encoders have long-term information propagation issues
(Jiao et al., 2019)	IEMOCAP Friends Emotion			/				/	/					/	Models' capability of understanding subtle emotions is still limited and more exploration is required.
(Xu et al., 2019)	IEMOCAP				/						/			/	Speech and text inherently co-exist in the temporal dimension

Table 2.4 : Continued

Author	Dataset	Input Representation				Approach								Remark
		GloVe	Word2Vec	BERT	OTHERS	CNN	GCN	GRU	BIGRU	LSTM	BILSTM	TRANSFORMER	MLP	ATTENTION	
(Tripathi et al., 2019)	IEMOCAP	/								/					LSTM is a common approach in sentiment analysis and NLP
(J. Li et al., 2020)	EmoryNLP MELD IEMOCAP			/								/	/		Model is able to effectively capture the whole context information of conversations and the speaker information
(J. Zhang et al., 2022)	IEMOCAP			/								/			BERT and Transformer show strong performance in natural language feature extraction.
(D. Sun et al., 2023)	IEMOCAP			/											The advent of BERT has revolutionized NLP by setting new performance standards across various tasks, establishing the "Pretrain + Finetune" paradigm as a standard approach.

Table 2-4 summarized the TER approach implemented in the previous studies. Many researchers in the field of text emotion recognition have increasingly emphasized the importance of learning context, with current approaches highlighting the remarkable success of models like BERT. BERT's ability to capture contextual information has indeed yielded very promising results and often outperformed other traditional methods in this domain. However, it is important to acknowledge that there are still some challenges in using BERT that need to be catered such as handling its computational intensity at the same time dealing with domain adaptation (Ganesh et al., 2021). This challenge remains an ongoing area of investigation with many experts in the field continuing their efforts to find effective solutions.

2.4.3 Facial Emotion Recognition (FER)

Facial expressions play a significant role in nonverbal communication and provide a lot of important information. Facial expressions are an important indication to observe a human’s emotions as it conveys various emotional states through the movement and action of facial features like eyes, eyebrows, nose, mouth, forehead and chin (Dinakaran & Ashokkrishna, 2020). Researchers are exploiting the combination of these facial features to improve recognition of emotion rates as facial expressions comprise various aspects that differ from person to person depending on gender, age, ethnicity, and other factors (Ghazouani, 2021). Recent technology developments provide various methods for autonomous facial emotion recognition (FER). Automatic facial emotion recognition is an intriguing study area that has been presented and used in a number of fields, including safety, health, education etc. There are many established databases on FER such as Extended Cohn–Kanade CK+ (Lucey

et al., 2010), Japanese Female Facial Expression (JAFEE) (Lyons, 2021), FER-2013 (Goodfellow et al., 2013) and many more.

There are two major procedures in a FER system which are feature extraction and classifier design. Handcrafted features were typically used in FER which were subsequently fed into various classifiers such as Support Vector Machine (SVM), Random Forest classifier (RFC), Linear Regression with Stochastic Average Gradient (SAG and many more. Numerous of geometric-based (position, angle, landmark points) and appearance-based (entire input image) feature extractors including Histogram of Oriented Gradients (HoG), Local Binary Pattern (LBP), Scale Invariant Feature Transform (SIFT), etc., have been continuously developed for the past few years. However, since the features are manually extracted and the procedure is laborious when the dataset is very large, these methods are unable to produce accurate and precise results (Chinchanikar, 2019).

Over the last ten years, Deep Learning (DL) approaches have grown significantly and they currently outperform traditional machine learning techniques in terms of accuracy and efficiency. With the development of DL approaches, features are now automatically extracted thus increasing the rate of recognition (D. Y. Choi & Song, 2020; Jung et al., 2015; Yang et al., 2018). Besides that, deep learning techniques are able to avoid the difficulty and complexity of manually extracting features. DL methods such as Convolutional Neural Network (CNN), Deep Belief Network (DBN) 3-Dimensional Convolutional Neural Network (3DCNN) and Recurrent Neural Network (RNN) have been chosen by a large number of researchers and developers.

CNN is a famous method of deep learning implemented by many researchers and previous works of FER. Study from (Lopes et al., 2017) proposed a simple approach for recognizing face expressions that combines CNN with specific image pre-processing techniques such as image cropping, rotation correction, spatial and intensity normalization. In the study from (Y. Li et al., 2018), they introduced the implementation of attention mechanism with CNN architecture for FER in the presence of occlusion images. Meanwhile, (Yolcu et al., 2019) proposed four cascades of CNN architecture in their study whereby the first three CNN are for segmentation of facial components while the fourth CNN is for the recognition of facial expressions. (Agrawal & Mittal, 2020) have investigated the effect of kernel sizes and number of filters of two versions of CNN architectures on the classification accuracy. Apart from that, (Jain et al., 2019) introduced residual blocks with the implementation of CNN. In general, all of the reviewed methods produced competitive results that exceeded 90% of accuracy in FER.

In most of the cases, the majority of previous works have employed the CNN for FER and it is proven to be effective when applied to data images. However, using CNN alone is insufficient to capture the temporal features in video data. Therefore, many different structures of DL integrating CNN with RNN have been proposed in prior works in order to cater spatio-temporal features for FER. For instance, RNN-LSTM has been proposed in the study from (D. H. Kim et al., 2017) to learn time scale-dependent information from facial expression sequences with varying numbers of frames. Meanwhile, (Liang et al., 2020) introduced BiLSTM network in their learnable end-to-end framework. Aside from that, (Yu et al., 2018) have proposed 3DCNN to extract spatio-temporal convolutional features from the image sequences

and Nested LSTM to jointly learn the multi-level appearance features and temporal dynamics of face expressions.

Table 2-5 summarized the FER approach implemented in the previous studies. Based on table, it is shown clearly that CNN are the dominant approach to FER, especially for static images and sequence images. However, one of the main gaps found in previous work on FER is that most of the work has focused on developing FER systems that are specifically designed for static images. This is likely due to the fact that static image FER datasets are more abundant and easier to collect than video FER datasets. Due to the limited availability of large-scale, high-quality facial expression video datasets, there is certainly a need for more techniques and approaches in which can help to improve the performance of FER by making them more robust to variations in the training data. Besides that, Video data is more complex than image data, and therefore, different techniques are needed for processing and extracting features from video data especially in challenging conditions such as detecting faces in videos when the face is not facing directly at the camera and recognize facial expression from limited information when the face is partially obscured.

Table 2.5 : Summary of FER in Previous Studies

Author	Dataset	Feature and Data Augmentation	Modality	Approach					Remark
				CNN	LSTM	3DCNN	BILSTM	RESIDUAL	
(Jung et al., 2015)	CK+ Oulu-CASIA	- Flip - Rotation	Image	/					Two deep network model cooperated to extract temporal from image sequences and trajectories of landmark points.
(Lopes et al., 2017)	CK+ JAFFE BU-3DFE	- Crop - Rotation - Intensity Normalization	Image	/					Special pre-processing steps are required due to lack of data in order to decrease the variations between images and to select a subset of the features to be learned
(D. H. Kim et al., 2017)	MMI CASMEII	-	Image	/	/				CNN and LSTM were able to learn spati0-temporal features in FER
(Yang et al., 2018)	BU-4DFE BP 4D CK+Oulu-CASIA MMI	-	Image	/					Learns the de-expression residue remained in the generative model. Capable of handling cases of both spontaneous expressions and posed expressions
(Y. Li et al., 2018)	CK+ Oulu-CASIA MMI	-	Image		/	/			3DCNN is used to extract spatio-temporal convolutional features from the image sequences that represent facial expressions while Nested LSTM is used to model the dynamics of expressions.

Table 2.5 : Continued

Author	Dataset	Feature and Data Augmentation	Modality	Approach					Remark
				CNN	LSTM	3DCNN	BILSTM	RESIDUAL	
(Yolcu et al., 2019)	RaFD	- Cascade	Image	/					Four-stage networks combines holistic facial information with local part-based features to achieve more robust facial expression recognition.
(Jain et al., 2019)	CK+ JAFFE	- Crop - Intensity - Normalization	Image	/				/	Residual able to overcome the challenge in computationally complex and lack of size and variety in dataset
(D. Y. Choi & Song, 2020)	MAHNOB-HCI INNA	- Resize	Image	/	/				Semi-supervised learning to reduce the annotation cost of FER-purpose dataset.
(Agrawal & Mittal, 2020)	FER-2013	- Rescale	Image	/					Kernel size and the number of filters have a significant impact on the accuracy of the network
(Liang et al., 2020)	CK+Oulu-CASIA MIMI	- Multitask - Cascade Convolutional	Image	/			/		Combination of dynamic CNN features and BILSTM memory excels in modelling the temporal information
(Tripathi et al., 2019)	IEMOCAP	-	Captured Data / Image	/	/				2DCNN offer a faster training and inference time
(J. Zhang et al., 2022)	IEMOCAP	-	Captured Data/ Image					/	Lack of technical power in motion capture leads to more noise in the data and it contains too few emotion feature

Table 2.5 : Continued

Author	Dataset	Feature and Data Augmentation	Modality	Approach					Remark
				CNN	LSTM	3DCNN	BILSTM	RESIDUAL	
(Akinduyite et al., 2024)	FER2013	- Crop - Histogram equalization	Image Video	/					The process comprises five stages: detecting the face within the image, identifying facial landmarks, extracting and tracking key facial features such as the eyes, eyebrows, and mouth, and finally classifying the facial expression

2.5 Summary of Literature Review

In summary, the literature review on emotion recognition provides a comprehensive exploration of both theoretical and technical aspects in the field, shedding light on the existing knowledge while also revealing critical limitations and gaps that drive research motivation in this intriguing field of study. The limitations, gaps, and research motivations can be highlighted as follows:

i. Limitation and Research Motivation in Multimodal Emotion Recognition:

- Limitation: Feature-level early fusion is preferred in multimodal emotion recognition, but it often encounters overfitting issues due to a large number of parameters, especially in trimodal systems.
- Gap: Research should explore alternative fusion approaches to address overfitting issues in feature-level early.
- Research Motivation: Investigating decision-level late fusion as an alternative and focus should be given in addressing loss fine-grain temporal spatial information issue and developing the best decision late fusion rule.

ii. Limitation and Research Motivation in Speech Emotion Recognition (SER):

- Limitation: Researchers have increasingly incorporated attention mechanisms and BILSTM approaches since 2019, but many studies still rely on spectral features, neglecting temporal context features that are crucial for capturing emotion dynamics in speech.
- Gap: There is a need to include temporal context features in SER as emotions exhibit temporal dynamics. Most prior studies inadequately address handling imbalanced class distributions, leading to overfitting and misclassification issues.

- Research Motivation: Research should focus on addressing the limitations by exploring methods to integrate temporal context features and handle imbalanced class more effectively.

iii. Limitation and Research Motivation in Text Emotion Recognition (TER)

- Limitation: BERT models have shown great success in capturing contextual information, but there are challenges in terms of computational intensity and domain adaptation when using BERT.
- Gap: Solutions to handle BERT's computational intensity and domain adaptation challenges are needed to make it more accessible and practical for text emotion recognition.
- Research Motivation: Find effective solutions to the challenge in BERT-based text emotion recognition models by focusing on developing a lightweight model and fine-tuning the model on a domain specific dataset.

iv. Limitation and Research Motivation in Facial Expression Recognition (FER):

- Limitation: Previous FER research has primarily focused on static images due to the abundance and ease of collecting static image datasets, leaving a gap in video FER research.
- Gap: There's a need for techniques and approaches that improve the performance of FER for video data, which is more complex and poses challenges like recognizing expressions from partially obscured faces or faces not directly facing the camera.
- Research Motivation: Researchers should work on developing FER systems tailored for video data, considering the challenges and complexities associated with video FER.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter will provide an overview of the research methodology used in this study, including the various steps and phases involved in the research process. Figure 3-1 presents the overview of the research framework involved in this study.

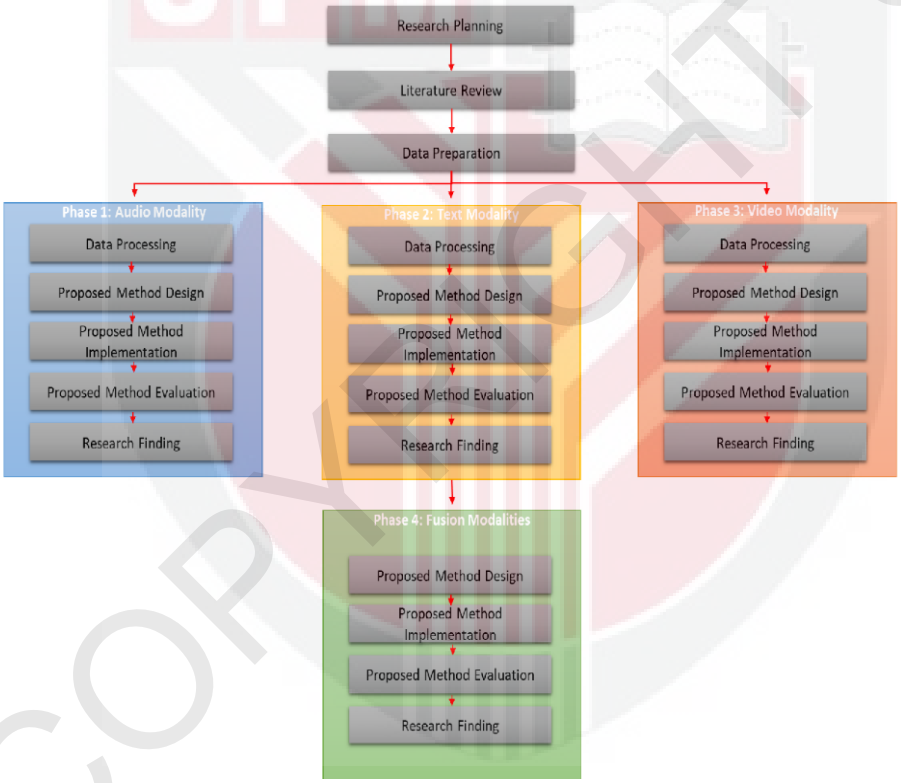

Figure 3.1 : Overview of the Research Framework

This study is organized into eight sequential steps which are Research Planning, Literature Review, Data Preparation, Data Processing, Proposed Method Design, Proposed Method Implementation, Proposed Method Evaluation, and Research

Findings. As outlined in Chapter 1, the primary objective of this research is to develop an accurate multimodal fusion model for emotion recognition.

To achieve this objective, the study is divided into four phases. The first three phases involve training individual models for each modality namely audio, text, and video. Each model is specifically designed and developed to achieve the best emotion recognition performance within its modality, aiming to produce accurate predictions. The fourth phase then integrates these results through a fusion model that combines the strengths of each modality.

The first three phases follow a systematic sequence of steps, beginning with Data Processing and continuing through Proposed Method Design, Proposed Method Implementation, Proposed Method Evaluation and concluding with Research Findings. The fourth phase, however, does not include a Data Processing step as it utilizes the outputs from the first three phases. Therefore, the fourth phase begins with Proposed Method Design and proceeds through the remaining steps.

This structured methodology allows for the comprehensive exploration of each modality's potential individually, setting a strong foundation for developing a robust and accurate multimodal fusion model for emotion recognition.

3.2 Research Planning

Research planning is the initial stage of any research project and is crucial to the success of the study. It is the process of determining what information needs to be gathered, how it will be implemented, and what resources will be required. A well-

planned research project can increase the chances of producing valid and reliable results. At this stage, this study starts with defining the problem statement towards the current issue relating to the field of multimodal emotion recognition. Following that, this study identifies the specific, measurable, and relevant research questions with regards to the research domain. From the research question, this study develops objectives in order to provide a clear understanding of what is the target and final goal to be achieved. Besides that, this study also defines the scope of study, selects the research design and outlines the data analysis strategy (refer Chapter 1).

3.3 Literature Review

Literature review is a critical component of research and involves systematically reviewing existing literature on a particular topic. The primary aim of literature review is to identify the gaps in knowledge and provide a conceptual framework for the study. It is also useful in identifying and evaluating different research methodologies and methods that have been used in previous studies.

To carry out a literature review, this study begins by defining the research questions or topic of interest. Then, this study searches for relevant literature from various sources such as journals, books, and other publications. One common approach used is electronic databases such as Google Scholar, PubMed, or Web of Science to search for articles, books, and other scholarly materials. This study uses various keywords and combinations of terms to optimize the search results. Besides that, this study also uses citation tracking to identify seminal works and trace the development of a particular research area.

After collecting the articles, each of the articles are read and evaluated carefully to determine their relevance and quality. The findings of the selected reviewed literature are synthesized by summarizing the key points, themes, and arguments made by various authors and identifying gaps. Finally, the literature reviews are organized and written in a clear and coherent manner to provide a comprehensive overview of the existing literature and how it relates to the research question (refer Chapter 2).

3.4 Data Preparation

Data preparation is a crucial stage in any research project that involves developing a deep learning model. The selection of a high-quality dataset is essential to ensure the accuracy and generalization of the model. Inaccurate or incomplete data can lead to biased results, which can compromise the validity and reliability of the research findings. To select the best dataset, this study typically reviews various published datasets and evaluates their suitability based on the research objectives and constraints. In addition, this study also searches for published research papers related to emotion recognition and checks the datasets used in the previous studies.

There are several factors to consider when selecting a published dataset. Firstly, the dataset should have a sufficient number of samples to represent the target population adequately. Secondly, the dataset should have appropriate labels that correspond to the emotions that are being recognized. Thirdly, the dataset should have a variety of emotional expressions to ensure that the model can recognize a broad range of emotions. Lastly, the dataset should be publicly available and have clear guidelines on how to use the data.

This study selects the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset, which meets all the above criteria. The IEMOCAP dataset is a widely used dataset for emotion recognition that contains audio and video recordings of actors in various emotional states. The dataset provides a vast range of emotional expressions and realistic interaction scenarios between speakers, making it suitable for developing a robust and generalizable emotion recognition model. Besides that, it has been extensively annotated, making it a suitable dataset for training and testing emotion recognition models. By selecting the IEMOCAP dataset, this study ensures that the emotion recognition model is trained on a diverse and high-quality dataset, enabling it to accurately recognize emotional expressions in various real-world scenarios.

3.4.1 IEMOCAP Dataset

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset will be used for training and evaluating the proposed system in this research study (Busso et al., 2008). IEMOCAP is an acted multimodal and multi-speaker database that was collected at SAIL lab, University of Southern California. This dataset contains 12 hours of audio-visual data, including video, speech, motion capture face and text transcription created by professional actors through a 5-minutes interaction. This dataset also comprises dyadic sessions in which actors execute improvised or scripted scenarios that are specifically selected to evoke emotional responses. Multiple annotators have annotated the IEMOCAP database with categorical labels such as anger, happiness, sadness, fear, surprise, neutrality as well as dimensional labels including valence, activation, and dominance.

3.4.2 IEMOCAP Data Processing

Data processing is an important stage in any research project that involves working with large datasets. This stage involves transforming raw data into a more usable format for analysis. The goal of data processing is to ensure the quality, consistency, and validity of the data to derive meaningful insights and draw accurate conclusions. In the case of the IEMOCAP dataset, which consists of three modalities: audio, text, and video image, the data processing stage is particularly important to ensure the accuracy and reliability of the results.

The dataset is organized into five sessions (Session 1 to Session 5), each containing subfolders for audio (sentences/wav), video (sentences/avi), and transcriptions (sentences/transcriptions). The emotion labels are located in the EmoEvaluation folder, where each session has a corresponding .txt file containing annotations. These annotations include the utterance ID (e.g., Ses01F_impro01_F000), start and end times, emotion labels (e.g., anger, happiness, sadness, neutral), and optional valence, arousal, and dominance scores.

The first step involves extracting and consolidating the emotion labels from all sessions into a single file named df_iemocap.csv. This file should include columns such as session, utterance_id, start_time, end_time, emotion, valence, arousal, dominance, and, later, text. Once the labels are prepared, the next task is to process the audio data. Using the start and end times from df_iemocap.csv, split the original WAV files into segments corresponding to each utterance, saving the results in a processed_audio folder.

Similarly, the video data in sentences/avi should be split into segments using the same timestamps. OpenCV is a suitable tool for video splitting, allowing the creation of individual video files for each utterance, which can be stored in a processed_video folder. For the textual modality, use the transcriptions provided in sentences/transcriptions to extract the text associated with each utterance. Match the timestamps and utterance IDs from df_iemocap.csv to retrieve the corresponding text and include it as a new column in the CSV file.

After completing the data processing, features from each modality will be extracted for further analysis. For the audio modality, MFCC (Mel-frequency cepstral coefficients), Delta and Delta Delta features will be computed to capture critical acoustic information. For the text modality, feature extraction will utilize BERT (Bidirectional Encoder Representations from Transformers) to generate contextualized embeddings. For the video modality, facial features will be detected using the Haar Cascade classifier, focusing on expressions and facial movements relevant to emotion recognition. The detailed methodologies and implementation of these feature extraction processes are explained comprehensively in Chapter 5, 6, and 7.

3.4.3 IEMOCAP Data Exploration

The IEMOCAP dataset is widely used in emotion recognition research but it is important to note that the distribution of emotional classes within the dataset is imbalance across the different modalities (audio, video, and text). Class distribution refers to the representation of each emotional category such as anger, happiness, sadness or neutral within the data. This imbalance can create challenges for machine

learning models, as they may become biased toward recognizing more frequent emotions thereby, underperforming for less common emotional categories. As a result, many researchers frequently use the most common emotions such as anger, happiness, sadness and neutral (Majumder et al., 2018; Poria, Chaturvedi, et al., 2016; Tripathi et al., 2019) . Following the previous work, this study will focus on improvising data with these four categories of emotions.

In the IEMOCAP dataset, the Happiness and Excitement categories are often merged to form a single Happiness category to address class imbalance as there are certain degree of resemblances inherent between both of them (Khalil et al., 2019). Building on the previous works trend, this study applies the same approach by merging these two categories to improve data representation. After merging, the combined Happiness category contains approximately 1636 samples. This makes Happiness one of the most dominant classes in the dataset, which further highlights the class imbalance when compared to less frequent emotions such as Sadness or Anger.

Overall, the IEMOCAP dataset demonstrates a heavily imbalanced class distribution across all its emotional categories. The most frequent emotion is Neutral, with 1,708 samples (31.7%), followed by Happiness, which accounts for 1,636 samples (30.4%). Anger appears in 1,103 samples (20.5%), and Sadness is the least frequent, with 1,084 samples (20.1%). This imbalance poses challenges for emotion recognition models, which often overfit to the dominant classes while struggling to accurately identify less represented emotions. Figure 1 illustrates the distribution of these emotions as percentages of the total dataset, highlighting the class imbalance that poses challenges for emotion recognition tasks.

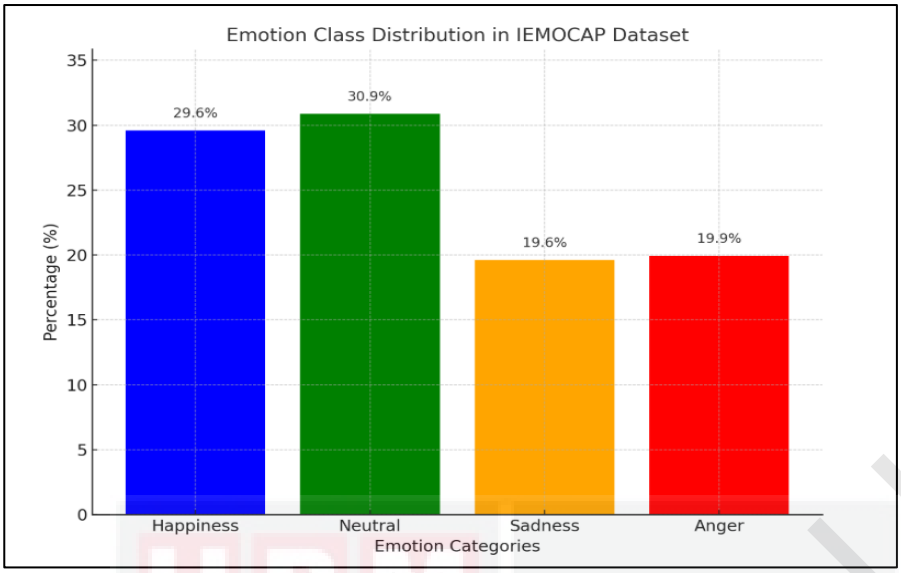

Figure 3.2 : IEMOCAP Data Distribution for Each Emotion

Table 3.1 : IEMOCAP Final Data Distribution

Categories of Emotion	Count
Neutral	1708
Angry	1103
Sad	1084
Happy	1636
Total	**5531**

3.5 Proposed Method Design

The proposed method design should involve careful consideration of various factors such as model design architecture, hyperparameter optimization, and model algorithm. Following is a breakdown of each of these components:

i. Model Design Architecture

- In this study, we introduced the Deep Hierarchical Weighted Late Fusion (DHWLF) model designed to integrate three modalities: audio, text, and video for robust emotion recognition. The model employs a late weighted fusion strategy, which combines the strengths of individual modality predictions through a weighted, hierarchical approach, ensuring that each

modality's unique contributions are effectively utilized. Unlike early fusion, which integrates raw data or features at the input level and can lead to challenges such as dimensionality issues and feature incompatibility, late fusion operates on the predictions of individual models. This allows each modality to be optimized independently with tailored architectures and preprocessing techniques, avoiding interference and ensuring better performance. The weighted late fusion approach is particularly advantageous as it dynamically accounts for variations in the reliability and relevance of different modalities in specific emotional contexts. By assigning higher weights to more reliable modalities, it reduces the influence of less effective ones, thereby enhancing the overall robustness, accuracy and adaptability of the model across diverse multimodal scenarios.

- This study introduced three individual models: SMOTE-2DCNN for audio, TextBERT for text, and Fer2DCNN for video to address the unique challenges associated with each modality. The design of these models was guided by the need to handle specific data-related issues effectively such as class imbalance in audio, contextual understanding in text and spatial feature extraction in video. These individual models serve as the foundation for the late fusion process, ensuring that the strengths of each modality are maximally leveraged. Each model was carefully implemented based on its ability to handle the unique characteristics of the input data and overcome limitations observed in alternative approaches. The selection of these models is strongly supported by findings from the literature, which highlight their superior performance and suitability for tasks similar to those addressed in this study.

A. Audio Modality

- SMOTE-2DCNN was employed to tackle two primary challenges: class imbalance and the need to capture spatial and temporal patterns in audio features such as Mel Frequency Cepstral Coefficients (MFCCs), delta and delta-delta coefficients. During preprocessing, SMOTE was applied to

synthesize new samples for underrepresented emotion classes, ensuring a balanced dataset and reducing the risk of biased predictions.

- SMOTE was specifically chosen over other oversampling methods such as random oversampling or Adaptive Synthetic Sampling (ADASYN) because of its ability to generate synthetic samples by interpolating between existing data points. **S**tudies such as (Chawla et al., 2002a) validate SMOTE's superiority in handling imbalanced datasets while maintaining the integrity of the feature space.

- After balancing the dataset, a 2D Convolutional Neural Network (2DCNN) was utilized to process the extracted features. The architecture comprised convolutional and pooling layers to capture hierarchical spatial patterns in the MFCCs, delta and delta-delta coefficients, followed by fully connected layers for classification. The use of 2DCNNs is supported by (T. Kim et al., 2019), which highlight their ability to exploit spatial correlations in structured audio features.

- SMOTE-2DCNN provides a robust and efficient solution. This combination, proven in the literature to yield significant improvements in audio-based emotion recognition, makes it an ideal choice for this study.

C. Text Modality

- TextBERT was implemented to address the challenges of text-based emotion recognition, which require a deep understanding of contextual relationships. The meaning of words in text can vary significantly depending on their surrounding words, and traditional models often struggle to capture such nuanced relationships.

- The implementation involved fine-tuning a pre-trained BERT (Bidirectional Encoder Representations from Transformers) model on the text data. Input text was first tokenized and converted into embeddings using BERT's transformer-based architecture. The final hidden states generated by the

transformer were then passed through a dense layer to perform emotion classification.

- BERT was selected for its bidirectional attention mechanism, which allows it to capture context from both preceding and succeeding words in a sentence. This capability makes it more effective than older models, such as Long Short-Term Memory (LSTM) networks, GloVe embeddings, or traditional bag-of-words approaches, which are often limited in their ability to understand context.

- The superiority of BERT has been established in studies like (Devlin et al., 2018) which demonstrate its state-of-the-art performance across a range of natural language processing tasks, including emotion recognition. Its ability to generalize well across datasets and adapt to domain-specific tasks through fine-tuning further solidified its selection as the text model in this study.

D. Video Modality

- The Fer2DCNN model was implemented to address the challenges of video-based emotion recognition, which relies heavily on accurate facial expression analysis. This requires models capable of effectively extracting spatial features such as edges, contours, and textures from video frames to classify emotions accurately.

- The implementation began by extracting video frames and detecting facial regions within each frame using a Haar cascade classifier. The detected facial regions were then passed to the Fer2DCNN model, which comprises convolutional layers designed to capture spatial patterns and dense layers for final emotion classification. The model was fine-tuned using facial emotion datasets to ensure it performed well across different emotion categories.

- The Haar cascades employ a hierarchical approach to detect faces by combining simple features in a cascading manner, ensuring precision and reliability. Studies such as (Viola & Jones, 2001) demonstrated the effectiveness of Haar cascades for real-time face detection with minimal false

positives, which makes them particularly suited for preprocessing in emotion recognition.

- The selection of Fer2DCNN was guided by its proven ability to extract meaningful spatial features for facial expression recognition. Compared to handcrafted methods like HOG+SVM or models like 3DCNNs, 2DCNNs strike a balance between accuracy and efficiency, excelling in frame-wise spatial feature extraction. Studies such as (Jain et al., 2019) and (Y. Li et al., 2018) have demonstrated the effectiveness of 2DCNNs in facial emotion recognition tasks, further validating their suitability for this study.
- The combination of the Haar cascade for accurate face detection and the Fer2DCNN model for robust emotion classification ensures a reliable and effective approach to video-based emotion recognition.

- Details about the design, figure illustration and implementation for the DHWLF model and each proposed individual model are presented in Chapter 4,5,6 and 7.

ii. Hyperparameter Optimization

- Hyperparameters are variables that determine the behavior of the model and affect the accuracy and performance of the model. These hyperparameters include the learning rate, batch size, number of epochs, and optimizer. The choice of hyperparameters can significantly affect the model's performance and must be optimized to obtain the best results.
- In this study, the optimal hyperparameter settings for the proposed method are determined through a trial-and-error approach. Each trial's results are tracked and used as feedback to determine the optimal combination of hyperparameters for model performance.

iii. Model Algorithm

- Model algorithm involves designing a set of instructions or rules that a computer can follow to perform a specific task. In general, the model

algorithm involves the following steps whereby the proposed model was trained using the training set, validated using the validation set and test using testing set.

Overall, the proposed method design should involve a thorough consideration of these components to ensure that the results of the model are accurate, scalable, and generalizable to new data.

3.6 Proposed Method Implementation

The proposed method implementation stage of research methodology involves the actual development and implementation of the proposed method. This part presents a brief overview of the requirements and the type of selections including programming language, libraries, Integrated Development Environment (IDE) and hardware used for the implementation of the proposed method.

Python is the main programming language used in this study for data processing and model development. This study uses the latest version of the Python programming language, Python 3 which was released in 2008. Python 3 is a high-level programming language that is easy to learn and use besides offering a vast array of libraries that are available for various purposes. Some of the basic libraries used in this study include NumPy, Pandas, Scikit-learn and Tensorflow. NumPy is used for scientific computing, Pandas for data manipulation, Scikit-learn for machine learning, and Tensorflow for deep learning. These libraries provide a rich set of functionalities such as data pre-processing, model training, and evaluation that can be leveraged to implement the proposed method efficiently.

For the development platform implement several IDE and web IDE such as Visual Studio Code, Jupyter Notebook and Google Colab. These IDEs provide a user-friendly interface for code development and debugging, as well as other features such as version control, code profiling, and code analysis. The proposed method involves training deep neural networks on a large dataset. Because of that reason, hardware requirements for the implementation of the proposed method requires a powerful Graphical processing unit (GPU) to speed up the training process. Therefore, this study chooses to use a dedicated GPU, such as NVIDIA's GeForce RTX and Tesla series provided in Google Colab for the training and testing of the model. Table 3-2 summarizes the requirements needed for the proposed method implementation.

Table 3.2 : Summary of the Requirement for Proposed Method Implementation

Requirement	Type of Selection
Programming Language	- Python 3
Library	- NumPy - Pandas - Scikit-learn - Tensorflow (Other related Libraries used are mentioned and explained in Chapter 4, 5, 6 and 7)
Integrated Development Environments (IDE)	- Visual Studio Code - Jupyter Notebook - Google Colab
Hardware	- Graphical processing units (GPU): ▪ Nivida GeForce RTX ▪ Tesla

3.7 Proposed Method Evaluation

In the proposed method evaluation stage, it is essential to assess the effectiveness and efficiency of the approach and algorithm in order to determine if it meets the research objectives and provides better results compared to existing methods. This stage

involves two important components which are experimental set up and performance metrics.

The first step is to set up experiments in order to evaluate the proposed method effectiveness and also compare its performance against other approaches from benchmark. To address potential inconsistencies in training and test segment selection across state-of-the-art studies, this study summarized the approaches used in the literature. Table 3-3 provides a comparative overview of the segmentation strategies adopted in previous research.

Table 3.3 : State-of-Art Training and Test Segmentation

Method	Training and Test Data Segmentation
BLSTM + SelfAttn (Santoso et al., 2022)	- Random Sampling. - Five-fold cross-validation. - The training set includes speech data from four sessions, and the test set contains data from the remaining session
CNN+RNN+3DCNN (M. Singh & Fang, 2020)	- Systematic Random sampling
MMFA-RNN (Ho et al., 2020)	- 10-fold - Leave-one-speaker-out cross validation method
CHFusion (Majumder et al., 2018)	- Systematic Random sampling
CMN (Hazarika, Poria, et al., 2018)	- Speaker-independent sampling
bc-LSTM + SVM (Poria et al., 2018)	- 5- fold - 80 % 20% training and validation set
BiLSTM -Attn +LSTM + CNN (Tripathi et al., 2019)	- Random Sampling - 77.7% Training - 20% Validation - 22.2% Testing

Following the insights gained from the comparison of training and testing strategies across state-of-the-art methods, this study adopts a segmentation approach tailored to

ensure fairness and consistency in evaluation. Specifically, this study carefully designs the division of the to align with best practices observed in the literature.

In all experiments of this study, the dataset is divided into training, validation, and testing sets, with each phase serving a distinct purpose. During the training phase, the model learns patterns and relationships within the data by minimizing error through iterative optimization. Features extracted from the dataset, such as MFCC, Delta and Delta Delta for audio, BERT embeddings for text and facial features for video are fed into individual models trained separately for each modality. In the validation phase, a separate validation set is used to fine-tune hyperparameters and monitor the performance of each model after each epoch to prevent overfitting. Metrics such as accuracy and F1-score are evaluated on the validation set to guide adjustments to model architectures or learning parameters. Finally, in the testing phase, the individual and fusion models are evaluated on a separate unseen dataset to provide a fair and unbiased assessment of their effectiveness in real-world scenarios.

This study used 80% of the dataset for training, 10% for validation, and 10% for testing by following a robust data-splitting strategy to ensure reliable evaluations. For that, stratified K-Fold Cross-Validation with a k value of 5 was applied to all individual and fusion model training to ensure balanced class representation across folds. The choice of k=5 for Stratified K-Fold Cross-Validation balances computational efficiency and evaluation robustness. A smaller k value, such as 2 or 3, may lead to fewer training samples in each fold, potentially reducing model performance consistency. Conversely, a larger k value such as 10, increases computational cost significantly without substantial improvement in validation reliability. Choosing 5

folds provides a practical compromise that offer sufficient training data in each fold while keeping the computational expense manageable. This ensures robust evaluation and reliable performance estimates for the individual models.

In the final stage of the study, the individual modality predictions were integrated using a Weighted Late Fusion approach to achieve multimodal emotion recognition. This approach aggregates the predictions from the audio, text, and video modalities at different hierarchical levels by assigning optimized weights to each modality based on its contribution to the final prediction. During the fusion phase, the training, validation, and testing processes were carried out on the combined predictions to fine-tune the weights and evaluate the overall performance. The stratified K-Fold method, with k=5 is also applied during this phase to optimize the fusion weights while ensuring balanced class representation. The testing phase in this stage assessed the final multimodal model's ability to generalize and effectively combine predictions from all modalities. The experiment setting detail will be further described in Chapter 4, 5, 6 and 7

Following the experimental set up, this study employ accuracy and Weighted F1 score as the evaluation metrics to assess the performance of the proposed model on the IEMOCAP dataset. These metrics are selected to provide a balanced and comprehensive evaluation of the model's effectiveness. IEMOCAP exhibits an imbalanced distribution of emotion classes. This necessitates evaluation metrics that can fairly represent the contributions of all classes regardless of their frequency.

Accuracy measures the overall proportion of correctly classified instances out of the total number of predictions. Aside that, the Weighted F1 score play the role as a complementary metric. It calculates the harmonic mean of precision and recall for each class and then computes a weighted average based on the support (number of true samples) of each class. This ensures that the performance of all classes, including minority ones, is considered proportionally, making it an ideal choice for the IEMOCAP dataset.

Unlike Micro F1 score, which aggregates results across all classes and can be dominated by majority classes, the Weighted F1 score gives fair representation to smaller classes without allowing them to disproportionately influence the overall score, as in the case of Macro F1 score. Therefore, the Weighted F1 score strikes an effective balance, making it well-suited for datasets with imbalanced class distributions like IEMOCAP.

In addition to accuracy and Weighted F1 score, this study computes and compares other metrics such as precision, recall, and per-class F1 score. These metrics provide a more granular view of the model's performance for individual emotion categories, ensuring a thorough analysis and enabling a fair comparison with several current state-of-the-art emotion recognition models.

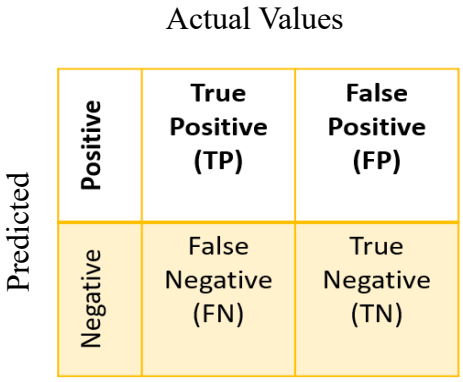

Figure 3.3 : Confusion Matrix Table

As illustrated in Figure 3-5, the confusion matrix is a square matrix with two rows and two columns, which are labelled as follows:

i. True Positive (TP): The number of instances that are correctly classified as positive.

ii. False Positive (FP): The number of instances that are incorrectly classified as positive.

iii. True Negative (TN): The number of instances that are correctly classified as negative.

iv. False Negative (FN): The number of instances that are incorrectly classified as negative.

By using these values, it can calculate the following performance metrics:

i. Precision

- Precision is a measure of the accuracy of positive predictions. It is calculated as the ratio of true positives to the sum of true positives and false positives.

$$\boldsymbol{Precision} = \frac{TP}{TP + FP}$$

(Equation 7)

ii. Recall

- Recall is a measure of the completeness of positive predictions. It is calculated as the ratio of true positives to the sum of true positives and false negatives.

$$\boldsymbol{Recall} = \frac{TP}{TP + FN}$$

(Equation 8)

iii. F1-score

- F1-score is a measure of the balance between precision and recall. It is calculated as the harmonic mean of precision and recall.

$$\boldsymbol{F1score} = \frac{2 * (precision * recall)}{(precision + recall)}$$

(Equation 9)

iv. Accuracy

- Accuracy is a measure of the overall performance of the classification model. It is calculated as the ratio of correct predictions to the total number of predictions.

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

(Equation 10)

v. Weighted F1-score

- The Weighted F1 score is the weighted average of these individual scores, where the weights are based on the number of true samples in each class.

- For a sample with *N* classes, the sample-weighted F1 score can be expressed as:

$$\boldsymbol{Weighted\ F1\ Score} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{w_i}\ \times F1score_i$$

(Equation 11)

where,

$$\boldsymbol{w_i} = \frac{Number\ of\ samples\ in\ class\ i}{Total\ number\ of\ samples}$$

3.8 Research Findings

The research findings indicate the final stage of the research framework. This stage represents the culmination of the research work where it presents and interprets the results of the study. This study should provide a clear and concise summary of the findings, interpret the results in light of the research questions and objectives, and reflect on the overall research process.

CHAPTER 4

DEEP HIERARCHICAL WEIGHTED LATE FUSION MODEL (DHWLF)

4.1 Introduction

This chapter introduces the Deep Hierarchical Weighted Late Fusion (DHWLF) model as the multimodal for emotion recognition. The DHWLF model employed the weighted average ensemble approach as the best rule for combining the three modalities, audio, text, and video image. Initially, the individual emotion recognition model of the three modalities were trained independently using the same dataset with unique features, representation, algorithm and hyperparameters. The DHWLF model then fused the predicted value of each individual emotion recognition model with different weighting factors based on their performance in order to obtain the final prediction. Further description regarding the proposed system and model will be presented in the following sections of this chapter.

4.2 Proposed Model of DHWLF

This section describes the proposed model of DHWLF for multimodal emotion recognition. In this proposed model, the DHWLF implements the weighted average ensemble approach to fuse the predictions of multiple individual emotion recognition models into a single prediction. This approach involves taking a weighted average of the individual emotion recognition model predictions, where each model is assigned a weight that reflects its relative importance.

4.2.1 DHWLF Design

The DHWLF model is designed in the form of a hierarchy to learn emotional knowledge across audio, text and visual domains and transmit the cross-domain knowledge through fusion. The design of the proposed DHWLF model is illustrated in Figure 4-1. It is divided into three deep learning models to train and classify three kinds of data modality; audio classification, text classification and video image classification. All the individual emotion recognition models will be trained independently using the data features extracted from the audio, text and video image. Each model has unique learning and hyperparameters tuned to the specific input form and situation. They will continue to be trained until achieving satisfactory accuracy results with no loss value reduction. After training, the models are evaluated on the testing dataset and the percentage of prediction accuracies are used to determine the optimal weighting factor of each model. Once the weighting factor is assigned, all the models will be finally fused by taking the weighted average in order to obtain the overall prediction.

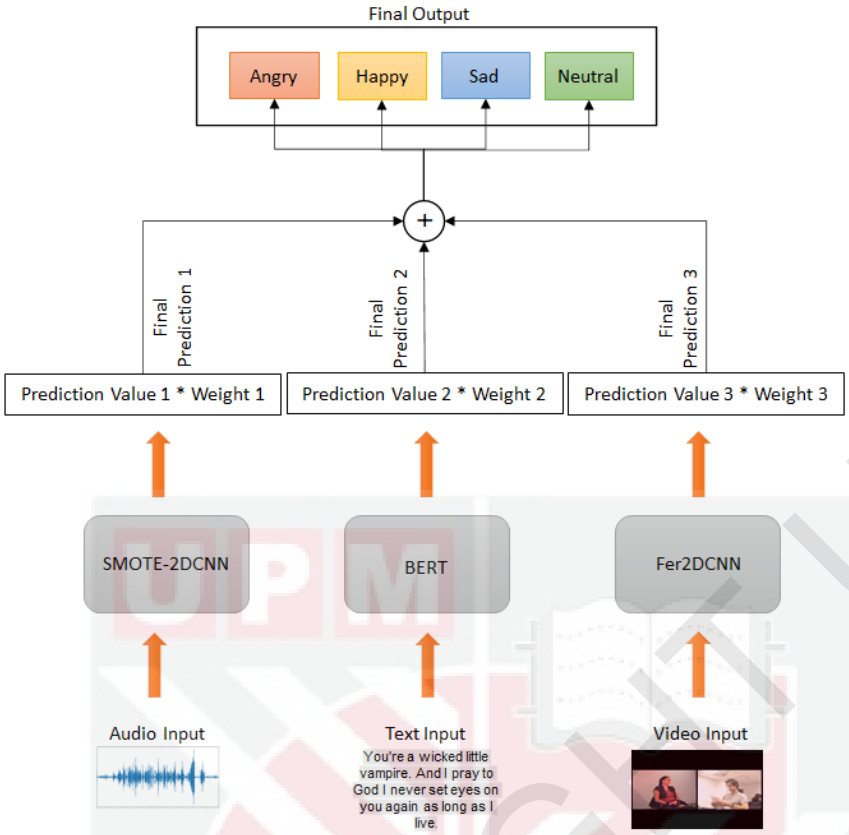

Figure 4.1 : Design of Propose DHWLF Model

4.2.2 Individual Emotion Recognition Model

This section presents the summary of the proposed individual emotion recognition model for audio, text and image data modality. Each individual model uses different numbers of training, validation, and testing samples due to the nature of the data in each modality and the preprocessing steps involved. The late fusion approach employed in this study accommodates differences in sample counts across modalities. For this reason, each modality is handled independently with variations such as filtering frames in video or excluding noisy audio, which lead to differences in the dataset sizes. Late fusion integrates the predictions from these independently trained models, ensuring flexibility and optimized performance for each modality. Further details on feature extraction, model design and algorithms, evaluation processes, and

additional results for each individual emotion recognition model are provided in Chapters 5, 6, and 7.

A. Audio Classification Model

This study introduced the SMOTE-2DCNN model as the audio classification model for emotion recognition. The MFCC, deltas, and deltas-deltas features extracted from the audio data signal of the IEMOCAP dataset served as the input to the classification model. The SMOTE-2DCNN model is composed based on two key modules which are imbalance handling module with SMOTE technique and classification module with 2DCNN architecture. The task of the imbalance handling module is to oversample the minority class in the dataset. This involved generating synthetic data points for the minority class using nearest neighbour algorithms, in order to balance the class distribution in the dataset.

Once the class distributions are balanced, the classifier module will perform the emotion label classification task. The 2DCNN consists of three convolutional layers and two fully connected layers. After each convolutional layer, max-pooling layer with a kernel size of two are used to reduce the data size, followed by a Batch Normalization

(BN) layer and an activation function ReLU. The proposed SMOTE-2DCNN managed to achieve 80% of accuracy with loss = 1.3.

B. Text Classification Model

This study implemented a BERT language model based on the transformer's encoder in the development of the text classification model known as TextBERT. The data

used for the text classification model is the readily accessible text transcription from the IEMOCAP dataset. The initial step consisted of preparing the input data using BERT. This encompassed the transformation of the input text into the format preferred by BERT. This format entails creating a sequence of tokens and incorporating special tokens at both the beginning and end. The special tokens include [CLS] (classification token) at the beginning and [SEP] (separator token) at the end of the text. Additionally, the input text was also tokenized using BERT's tokenizer, which possesses its individual vocabulary.

After preparing the input data, the input data was sent into the proposed TextBERT model for fine-tuning. Beginning with a pre-trained TextBERT model, a specific fine-tuning process was applied for the given text classification task. Alongside the TextBERT architecture, a fine-tuning model was developed by incorporating a dense layer. This layer utilized the contextual encoding representations generated by TextBERT to execute the classification task. The fine-tuning process included multiple rounds of training, adjusting and configuring the pertinent hyperparameters like learning rate and batch size for the purpose of optimizing the model's performance. After the training was completed, the model's performance was evaluated using the test dataset and the text classification model achieved 71% accuracy with loss = 2.5.

C. Video Classification Model

This study introduced Fer2DCNN for the video classification model. Fer2DCNN is a customized 2DCNN algorithm with an advanced pre-processing method that focuses on facial expression detection. Even though video data from the IEMOCAP dataset

was employed, noticeable distinctions between successive frames in the motions and facial expressions were not abundant. Therefore, rather than directing attention to temporal features, the choice was made to emphasize spatial features to mitigate the risk of overfitting and enhance data dimensionality. From the video, individual video frames were extracted and processed as single images. The Cascade Haar Classifier approach was subsequently applied for face detection following the extraction of frames from the video data.

The 2DCNN model was constructed utilizing two convolutional layers and two fully connected layers. The input shape of the 2DCNN model is 64 x 64 x 3. After each convolutional layer, a max-pooling layer with a kernel size of two are used to reduce the data size, followed by an activation function, ReLU. The proposed model was trained using the Adam optimizer, employing a default learning rate of 0.01. The Fer2DCNN architecture achieved an accuracy of 78% and a loss of 0.85.

4.2.3 Determination of Optimal Weighting Factor

The weighting factors determine the importance of each model in making the final prediction. Following the training of each individual emotion recognition model, the assignment of weights is necessary to reflect the accuracy of predictions made by each model. This involves determining the optimal weighting factor for each model. The calculation of the weighting factor for each model can be accomplished using the subsequent formula:

For the i^{th} model, let acc_i be its percentage accuracy. Then, the weighting factor, w_i can be calculated as:

$$w_i = \frac{acc_i}{\sum_{i=1}^{nm} acc_i}$$

(Equation 12)

where, *nm* is the total number of models. The denominator in the formula represents the sum of all the accuracies of the models. By dividing the accuracy of the i^{th} model by this sum, a weight is obtained that reflects the relative accuracy of that model.

4.2.4 Weighted Late Fusion of Models

Once the weighting factors for each model are determined, a weighted average can be employed to derive the overall prediction of the fusion model. Let P_i represent the prediction of the i^{th} model, alongside w_i as the associated weighting factor. Subsequently, the ultimate prediction denoted as P_{final}, can be computed as follows:

$$P_{final} = \sum_{i=1}^{n} w_i * P_i$$

(Equation 13)

This formula takes the weighted predictions of each model, where the weight of each model is given by its weighting factor. Finally, the following formula is employed to derive the predicted label class on the provided sample, $FinalOutput_T$, representing the ultimate output from the proposed model:

$$FinalOutput_{Ti} = \max \left[P_{final}\right]_1^{nc}$$

(Equation 14)

where, T_i is the given input sample and *nc* is the number of classes. Furthermore, to obtain the final percentage accuracy, acc_{DHWLF} of the proposed DHWLF model is

determined by computing the ratio of correctly predicted samples to the total number of samples in the dataset and multiplying this ratio by 100 to express it as a percentage.

4.2.5 DHWLF Algorithm

The DHWLF model's proposed set of steps which is outlined in Algorithm 1. Let D be the set of data, where: $D = \{(x_A, y_A), (x_T, y_T), (x_I, y_I)\}$. Here, (x_A, y_A), represents the feature set and label of audio data A, (x_T, y_T), represents the feature set and class label of text data T, and (x_V, y_V), represents the feature set and label of video data V. Each modality A, T, and V corresponds to a different type of data used in the model, and the dataset is prepared and divided into three segments: training, validation, and testing.

The training dataset D is utilized to sequentially train a set of n models, where $n = 3$, corresponding to the three modalities. Note that each modality uses different numbers of training, validation, and testing samples. For each individual model Mi, where i corresponds to each model's index, the input data D_j corresponds to the feature set and label of the j^{th} modality, where $j \in \{1,2,3\}$, representing audio, text, and video, respectively.

Following the training process, each model is used to provide predictions on the validation data. These predictions are then utilized to compute the accuracy acc_i for each model Mi. The weight factor w_i for each model Mi is determined based on its accuracy acc_i. Subsequently, the final output of the proposed DHWLF model is calculated by applying Equation 13, which involves multiplying the prediction value P_i from each individual emotion recognition model with the corresponding weight factor w_i . The sum of all these weighted predictions from each model yields the final

prediction P_{final}. To determine the predicted class for a given sample, the argmax function is applied to the output of the proposed DHWLF model, as described in Equation 14.

Algorithm 1: DHWLF Fusion model	
Input:	$D = \{(x_A, y_A), (x_T, y_T), (x_V, y_V)\}$ Split into Training, Validation, Testing
Output:	P_{final} = Final prediction $FinalOutput_{Ti}$ = Predicted label class on provided sample, T_i acc_{DHWLF} = Percentage accuracy of DHWLF model
1) For $i = 1 \ldots nm$, where nm is the number of models	a) For each modality, j, train model M_i using data Dj b) Get predictions value of M_i c) Compute the accuracy acc_i of model M_i
2) For $i = 1 \ldots nm$, where nm is the number of models	a) Calculate weight factor w_i of M_i using acc_i
3) Calculate final prediction of proposed DHWLF model, P_{final}	
4) Compute the output of proposed DHWLF model for each Ti, $i = 1 \ldots nc$ to get the predicted label class, $FinalOutput_{Ti.}$, where T_i is the given input sample and nc is the number of classes	
5) Get the final percentage accuracy of proposed DHWLF model, acc_{DHWLF}	
6) End	

Let's illustrate the flow of Algorithm 1 with a simple example:

i. Suppose there are 3 modalities: Audio, Text, and Video (denoted as D_1, D_2, D_3 respectively).

ii. Suppose have 3 models: M_1 for Audio, M_2 for Text, and M_3 for Video.

The steps would look like this:

1. For $i = 1 \ldots 3$ (models M_1, M_2, M_3):
    - For $i = 1$ (Model M_1 for Audio):
        - Train M_1 using D_1 (audio data).

        - Get predictions of M_1.
        - Compute accuracy acc_1 of M_1.
    - For i = 2 (Model M_2 for Text):
        - Train M_2 using D_2 (text data).
        - Get predictions of M_2.
        - Compute accuracy acc_2 of M_2.
    - For i = 3 (Model M_3 for Video):
        - Train M_3 using D_3 (video data).
        - Get predictions of M_3.
        - Compute accuracy acc_3 of M_3.
2. For i = 1…3 (computing weights for each model):
    - For i = 1 (Model M_1 for Audio):
        - Compute weight w_1 based on acc_1.
    - For i = 2 (Model M_2 for Text):
        - Compute weight w_2 based on acc_2.
    - For i = 3 (Model M_3 for Video):
        - Compute weight w_3 based on acc_3.
3. Calculate the final prediction of the proposed DHWLF model by combining weighted predictions from all modalities.

$$P_{final} = w_1 \times P_1 + w_2 \times P_2 + w_3 \times P_3$$

4. Determine the Final Class:
    - Apply the argmax function to the weighted predictions:
        - If the weighted sum favors Class 1, the final output will be Class 1.
5. Compute final accuracy acc_{DHWLF} of the DHWLF model.

4.3 Evaluation and Result

This section will discuss the evaluation process of the proposed method including the experimental design and the results obtained. The dataset used to evaluate the proposed model is the IEMOCAP dataset. The emotions of angry, happy, sad and neutral were selected from all categorical emotion labels annotated in IEMOCAP to maintain consistency with the previous research.

4.3.1 Experimental Set Up for DHWLF Model

This study proposed a DHWLF model that combines information from multiple modalities of data such as audio, text, and images to achieve better emotion recognition performance. To evaluate the effectiveness of the proposed model, a comprehensive evaluation setup has been devised, encompassing three experimental designs.

The first experimental design involves comparing the performance of the fusion model with the performance of individual emotion recognition models. By comparing the performance of these models, the contribution of each modality to the overall performance of the model can be determined and an assessment can be made regarding whether the fusion process leads to improved performance. This analysis also will help to identify the strengths and weaknesses of the model in recognizing emotions from different modalities of data.

The second experimental design aims to compare the performance of the weighted late fusion model with early fusion model. The early fusion method, based on the approach described in Tripathi et al.'s, 2019 study, integrates information from multiple

modalities at an early stage before performing the classification task, whereas the weighted late fusion model combines the outputs of individual emotion recognition-specific models at a later stage using a weighting factor. This experiment used the same method for the classification of individual modalities. The only difference is in the fusion approach either between the late fusion or early fusion. By comparing the performance of these two approaches of fusion models, it is possible to assess the impact of the fusion strategy on the overall performance of the emotion recognition model.

The third experimental design involves comparing the performance of the proposed method with other state-of-the-art methods in multimodal emotion recognition. The model is evaluated on IEMOCAP datasets, and its performance is compared with state-of-the-art models in the field of emotion recognition. This comparison helps in evaluating the effectiveness of the model and identifying areas for potential enhancement.

4.3.2 Result

The results and comparative analysis from the evaluation of the proposed DHWLF model based on the aforementioned experimental design are presented in this section.

A. Experiment 1: Comparison with Individual Emotion Recognition Models

The comparison of the proposed method with other individual emotion recognition models trained on the same dataset is presented in Table 4-1 and Figure 4-2, which also provides a graphical representation. From Table 4-1 and Figure 4-2, the image classification model (Fer2DCNN) performed best among the individual emotion

recognition models, achieving an overall accuracy of 84%. Meanwhile, the audio (SMOTE-2DCNN) and text (TextBERT) classification models yielded overall accuracies of 81% and 70%, respectively. Nevertheless, the proposed multimodal fusion model (DHWLF) achieves a significantly higher overall accuracy of 90%, demonstrating its ability to effectively integrate information from multiple modalities. This represents an improvement of 9% over the audio classification model, 20% over the text classification model and 6% over the image classification model, highlighting the robustness of the fusion approach in addressing modality-specific limitations.

Table 4.1 : Percentage Accuracy Comparison of Individual Emotion Recognition Model and DHWLF

Model	Modality	Overall Accuracy (%)
SMOTE-2DCNN	Audio	81
TextBERT	Text	70
Fer2DCNN	Image	84
DHWLF (Proposed Method)	Audio Text Image	**90**

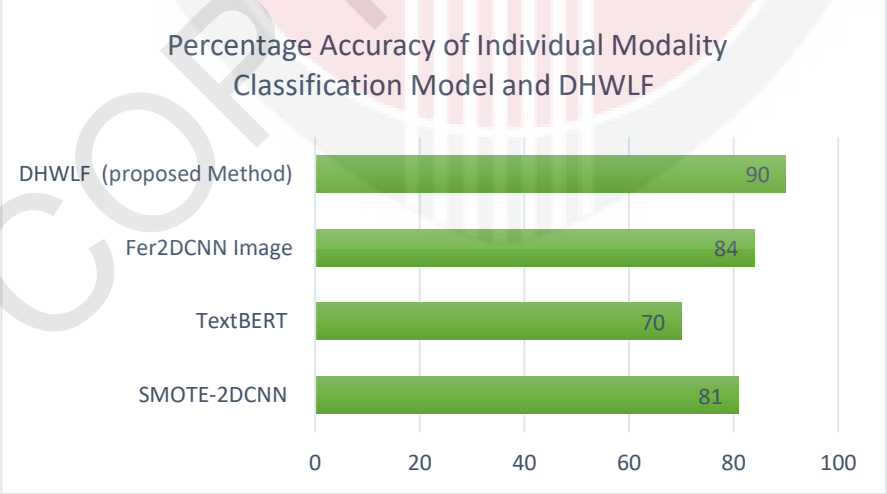

Figure 4.2 : Percentage Accuracy of Individual Emotion Recognition Model and DHWLF

Table 4-2 and Figure 4-3 present the comparison of the True Positive Rate (TPR) for each label class, with graphical representation for both the individual emotion recognition models and the proposed multimodal fusion model. Based on Table 4-2 and Figure 4-3, the proposed method achieves the highest TPR for two out of four label classes, demonstrating its effectiveness in handling diverse emotional expressions. Notably, it correctly identifies 100% of the samples belonging to the angry and neutral label classes, showcasing its superior ability to classify these emotions with perfect accuracy. Meanwhile, the audio classification model and image classification model perform best for the happy and sad label classes, achieving TPRs of 92% and 89%, respectively.

Table 4.2 : True Positive Rate Comparison of Each Class for Individual Emotion Recognition Model and DHWLF

Model	Modality	Angry (%)	Happy (%)	Sad (%)	Neutral (%)
SMOTE-2DCNN	Audio	83	91	**89**	61
TextBERT	Text	81	74	73	54
Fer2DCNN Image	Image	67	**92**	78	100
DHWLF (Proposed Method)	Audio Text Image	**100**	80	80	**100**

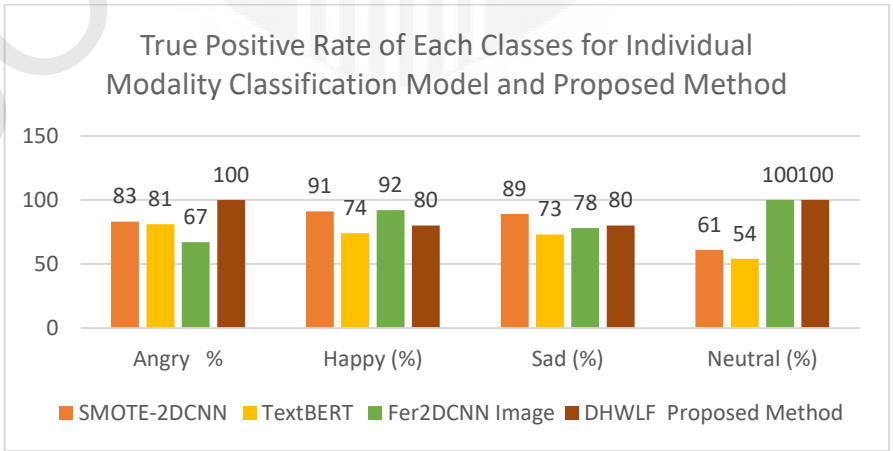

Figure 4.3 : True Positive Rate of Each Classes for Individual Emotion Recognition Model and DHWLF

Table 4-3 illustrates the performance comparison of individual emotion recognition models and the proposed method of multimodal fusion model. Looking at the table, it shows that the proposed method of multimodal fusion model has the highest precision and F1-score for all four label classes. However, for the recall result, audio classification model and image classification model yield the highest for sad and happy label class with 0.89 and 0.92, respectively. Despite the observed results, the proposed method remains effective in delivering competitive performance across all metrics. The slight variations in Precision, Recall, and F1-Score for specific labels reflect the complexities of the dataset and the multimodal fusion process rather than any fundamental limitations.

Table 4.3 : Performance Comparison of Individual Emotion Recognition Model and DHWLF

Performance		**Audio SMOTE-2DCNN**	**Text TextBERT**	**Image Fer2DCNN**	**Audio +Text +Image (DHWLF Proposed Method)**
Precision	Angry	0.78	0.69	**0.80**	**0.83**
	Happy	0.89	0.76	0.92	**1.00**
	Sad	0.82	0.70	0.70	**1.00**
	Neutral	0.72	0.65	**1.00**	0.83
Recall	Angry	0.83	0.81	0.67	**1.00**
	Happy	0.91	0.74	**0.92**	0.80
	Sad	**0.89**	0.73	0.78	0.80
	Neutral	0.61	0.54	**1.00**	**1.00**
F1-Score	Angry	0.81	0.74	0.73	**0.91**
	Happy	**0.90**	0.75	0.92	0.89
	Sad	0.85	0.71	0.74	**0.89**
	Neutral	0.66	0.59	**1.00**	0.91

B. Experiment 2: Comparison with Early-Fusion Approach

The comparison of the late fusion approach proposed method with the early fusion approach method, both trained on the same dataset, is presented in Table 4-4 and Figure 4-4, which also provides a graphical representation. From Table 4-4 and Figure 4-4, the late fusion approach proposed method outperformed the early fusion approach, achieving an overall percentage accuracy of 90% compared to 62% for the early fusion approach. This significant improvement of 28% highlights the effectiveness of the late fusion strategy in leveraging the strengths of individual modalities more efficiently, addressing limitations associated with early integration. The results underscore the robustness of the proposed method in capturing complementary information across modalities to achieve superior performance.

Table 4.4 : Percentage Accuracy Comparison of Early Fusion Model and Late Fusion Model (Proposed Method)

Fusion Approach	Modality	Overall Accuracy (%)
Early Fusion (Tripathi et al., 2019)	Audio Text Image	62
Late Fusion (Proposed Method)	Audio Text Image	**90**

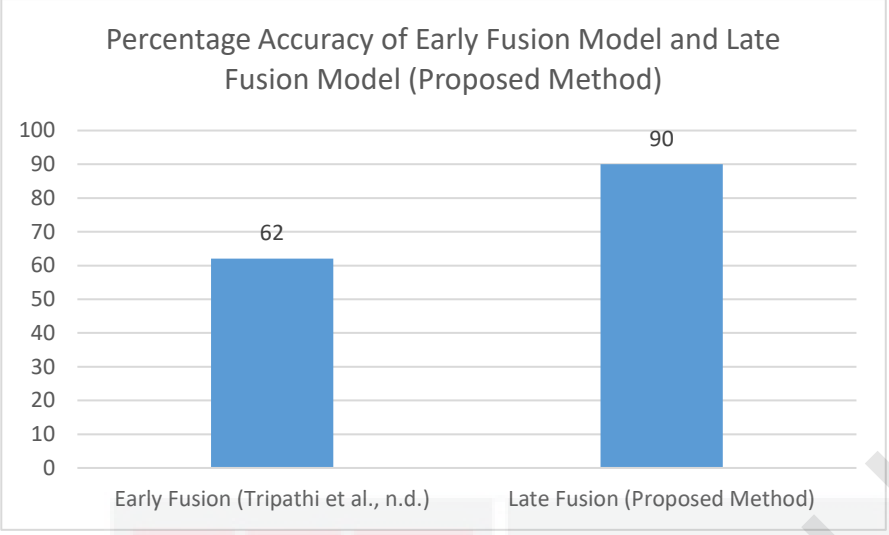

Figure 4.4 : Percentage Accuracy of Early Fusion Model and Late Fusion Model (Proposed Method)

Table 4-5 and Figure 4-5 present the comparison of the True Positive Rate (TPR) for each label class, with graphical representation for the late fusion approach proposed method and the early fusion approach method. Based on Table 4-5 and Figure 4-5, the late fusion approach proposed method achieves the highest TPR for all four label classes, demonstrating its superior capability in accurately recognizing emotions across diverse categories. Specifically, it correctly identifies 100% of the samples belonging to the angry and neutral label classes and 80% of the samples for the happy and sad label classes. These results highlight the robustness of the late fusion approach in consistently outperforming the early fusion method, particularly in handling emotions with varying levels of complexity.

Table 4.5 : True Positive Rate Comparison of Each Classes for Early Fusion Model and Late Fusion Model (Proposed Method)

Fusion Approach	Modality	Angry (%)	Happy (%)	Sad (%)	Neutral (%)
Early Fusion (Tripathi et al., 2019)	Audio Text Image	75	40	63	63
Late Fusion (Proposed Method)	Audio Text Image	**100**	**80**	**80**	**100**

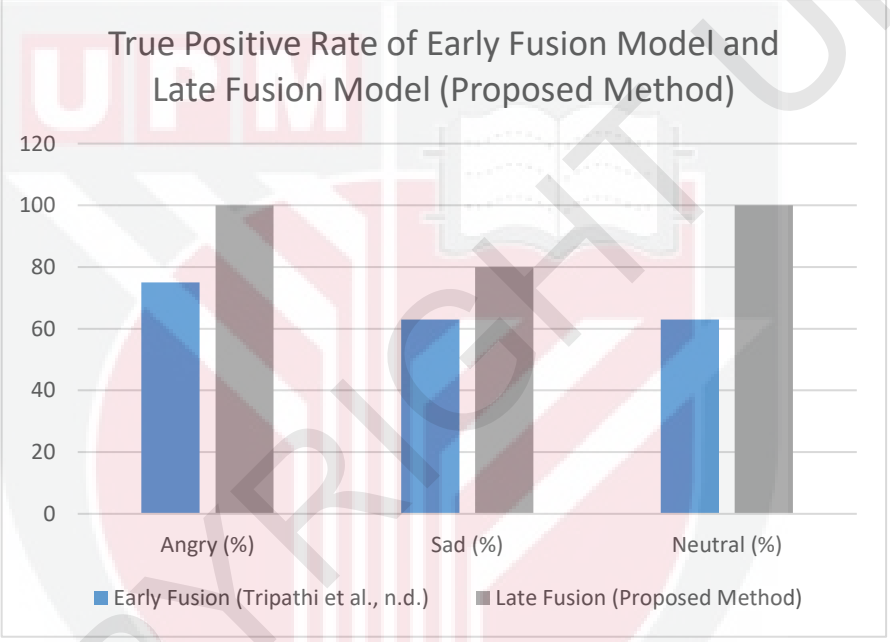

Figure 4.5 : True Positive Rate Comparison of Each Classes for Early Fusion Model and Late Fusion Model (Proposed Method)

Table 4-6 presents the performance comparison between the early fusion approach and the late fusion approach methods. According to Table 4-6, the late fusion approach proposed method achieves the highest precision, recall, and F1-score for all four label classes, highlighting its superior ability to balance accuracy and robustness across various metrics. These results demonstrate that the late fusion approach not only enhances detection accuracy but also ensures consistent and reliable performance in classifying diverse emotional labels compared to the early fusion approach.

Table 4.6 : Performance Comparison of Early Fusion Model and Late Fusion Model (Proposed Method)

Performance		Early Fusion	Late Fusion (Proposed Method)
Precision	Angry	0.59	**0.83**
	Happy	0.60	**1.00**
	Sad	0.59	**1.00**
	Neutral	0.67	**0.83**
Recall	Angry	0.75	**1.00**
	Happy	0.33	**0.80**
	Sad	0.63	**0.80**
	Neutral	0.63	**1.00**
F1-Score	Angry	0.66	**0.91**
	Happy	0.43	**0.89**
	Sad	0.61	**0.89**
	Neutral	0.65	**0.91**

C. Experiment 3: Comparison with State-of-The-Art

Table 4-7 and Figure 4-6 present the comparison of percentage accuracy with a graphical representation for state-of-the-art methods and the proposed method in multimodal emotion recognition. The presented state-of-the-art methods have taken into account the involvement of four emotions of the same label classes, namely angry, happy, sad, and neutral, trained on the same dataset. The unweighted accuracy (UA) and weighted accuracy (WA) were considered to ensure consistency with prior studies of the state-of-the-art methods. From Table 4-7, it can be seen that some of the state-of-the-art methods combine different modalities, such as Audio + Text, Audio + Visual, Audio + Text, and Audio + Text + Visual, in their multimodal emotion recognition approaches. Notably, the visual modality varies across studies, referring to visual stimuli like video, image, or motion capture data. On a scale of 70% to 78%, the UA and WA results of most state-of-the-art methods are relatively similar, regardless of the modality combinations.

Nevertheless, the proposed multimodal fusion model achieves the highest UA and WA, with both metrics reaching 90%. This represents a significant improvement of approximately 15% over the best-performing state-of-the-art methods. These results underscore the effectiveness of the proposed method in leveraging complementary multimodal data, surpassing the limitations of prior approaches and setting a new benchmark in multimodal emotion recognition accuracy.

Table 4.7 : Percentage Accuracy Comparison of State-of-The-Art and DHWLF

Method	Modality	UA (%)	WA (%)
BLSTM + SelfAttn (Santoso et al., 2022)	Audio + Text	76.8	76.6
CHFusion (Majumder et al., 2018)	Audio + Text	76.1	-
CHFusion (Majumder et al., 2018)	Audio + Visual	69.5	-
CHFusion (Majumder et al., 2018)	Text + Visual	75.9	-
CNN+RNN+3DCNN (M. Singh & Fang, 2020)	Audio + Visual	71.8	-
MMFA-RNN (Ho et al., 2020)	Audio + Text	73.2	-
CHFusion (Majumder et al., 2018)	Audio + Text + Visual	76.5	-
CMN (Hazarika, Poria, et al., 2018)	Audio + Text + Visual	-	77.6
bc-LSTM + SVM (Poria et al., 2018)	Audio + Text + Visual	71.6	-
BiLSTM -Attn +LSTM + CNN (Tripathi et al., 2019)	Audio + Text + Visual	67.3	71.0
MAFN (J. Zhang et al., 2022)	Audio + Text + Visual	71.0	75.6
DHWLF (Proposed Method)	Audio + Text + Visual	**90.0**	**90.0**

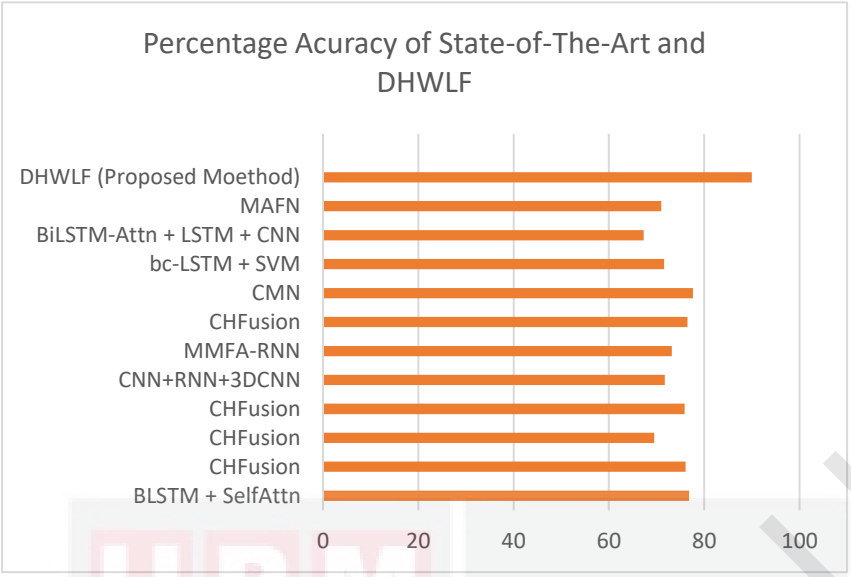

Figure 4.6 : Percentage Accuracy of State-of-The-Art and DHWLF

Table 4-8 presents the comparison of the percentage F1-score for each emotion label class, with a graphical representation for state-of-the-art methods and the proposed method in multimodal emotion recognition. In line with previous studies, the percentage F1-score was highlighted to provide a balanced measure of precision and recall, which is particularly valuable when addressing imbalanced datasets. Based on Table 4-8 and Figure 4-7, the proposed method outperforms the other state-of-the-art methods, achieving an average F1-score of 90% and yielding the highest F1-score across all four label classes. These results demonstrate the robustness of the proposed method in consistently identifying emotions with high precision and recall, underscoring its ability to handle imbalanced data effectively while outperforming existing approaches in multimodal emotion recognition.

Table 4.8 : Percentage F1 Score Comparison of State-of-The-Art and DHWLF

Method	Modality	F1-Score (%)				Avg. F1-Score (%)
		Angry	Happy	Sad	Neutral	
BLSTM + SelfAttn (Santoso et al., 2022)	Audio + Text	80.6	83.0	77.6	66.2	76.85
CHFusion (Majumder et al., 2018)	Audio + Text + Visual	77.6	81.4	77.0	71.2	76.80
MTA (Xi et al., 2023)	Audio + Text + Visual	87.19	84.47	82.78	68.07	80.63
DHWLF (Proposed Method)	Audio + Text + Visual	**91.0**	**89.0**	**89.0**	**91.0**	**90.00**

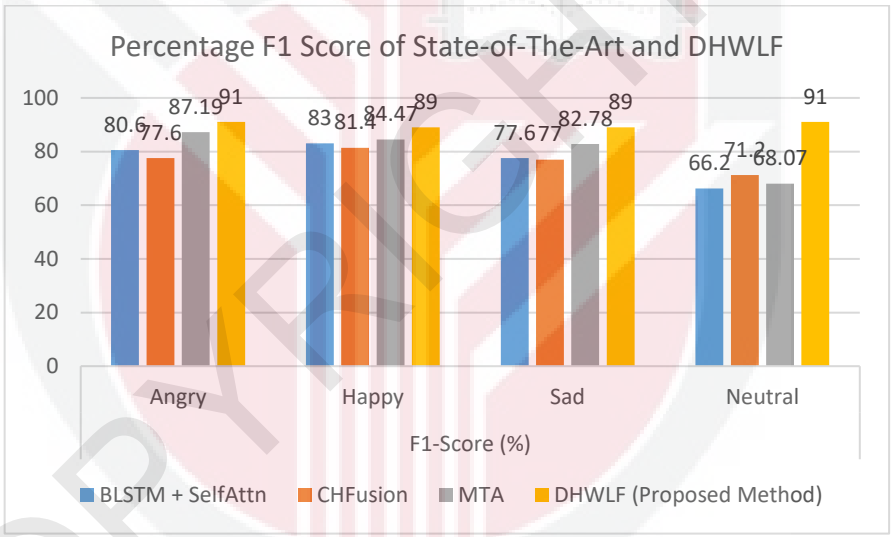

Figure 4.7 : Percentage F1 Score of State-of-The-Art and DHWLF

4.3.3 Analysis and Discussion

This section discusses the results for the different experimental settings mentioned in section 4.3.1.

i. *Experiment 1:* The first experiment compared the performance of the deep hierarchical weighted late fusion model (DHWLF) with the three individual emotion recognition models: SMOTE-2DCNN (audio), BERT (text), and Fer2DCNN (image), whereby the result shows that the DHWLF

fusion model outperformed the other individual emotion recognition models. There are several reasons why the DHWLF fusion model outperforms the individual emotion recognition models. Firstly, the fusion of multiple modalities allows for a more comprehensive understanding of the emotional state of the person being analyzed. For example, facial expressions from the image modality may reveal the intensity of the emotion, while speech patterns from the audio modality may indicate the specific type of emotion being experienced. Moreover, the integration of multiple modalities provides a more robust and reliable system that can cope with the inherent variability in individual modalities. The individual emotion recognition models may suffer from overfitting or underfitting, which can lead to inaccurate predictions. However, with the weighted late fusion approach, it can mitigate these issues by taking into account the strengths and weaknesses of each modality and generating a more robust prediction.

The performance of the proposed method on the Happy and Sad labels, as shown in Table 4-2 reflects an interesting trend. While these two labels perform exceptionally well in image and audio classification models, their performance in the proposed method is comparatively lower in this experiment. However, this can be attributed to specific characteristics of these labels and the fusion process rather than a limitation of the model. One possible reason is the modality-specific variations of the Happy and Sad labels. In image and audio modalities, these emotions are often associated with clear and distinct visual and auditory patterns such as smiling expressions or high-pitched tones for Happy and frowning expressions or low-pitched tones for Sad. These strong modality-specific features allow unimodal models to excel at identifying these emotions. In contrast, the fusion model integrates information from multiple modalities, which may lead to subtle variations in the decision-making process for these two labels due to the aggregation of features across all modalities. Importantly, the fusion model still demonstrates good performance on Happy and Sad, as evidenced by its ability to leverage complementary information from multiple modalities effectively. The results show that the

fusion model outperforms unimodal models overall by balancing strengths from all modalities, which highlights its robustness and effectiveness in handling multimodal emotion recognition tasks. Thus, the proposed method offers a balanced and comprehensive approach to emotion recognition, maintaining strong results for Happy and Sad while delivering superior overall performance across all labels. This underscores the fusion model's ability to harmonize and utilize multimodal information effectively, even for labels that excel in specific unimodal settings.

Apart from that, in terms of performance from the individual modalities, the result shows that the image modality achieved the highest accuracy among the individual emotion recognition models, followed by audio and text. This finding is consistent with previous studies that have found that facial expressions are particularly effective in conveying emotions, while vocal cues such as tone of voice and pitch are also important (Van den Stock et al., 2007). Text, on the other hand, may not be as effective in conveying emotions due to the lack of nonverbal cues.

ii. *Experiment 2:* The second experiment compared the performance of the weighted late multimodal fusion model with early fusion model. Overall, results demonstrate that the weighted late multimodal fusion model performed far exceeding the early fusion model.

The superiority of the weighted late fusion model over the early fusion model can be attributed to several factors. Firstly, the individual modalities may have different levels of relevance or importance in predicting the target emotion, and with the weighted late fusion approach, it can take this into account by assigning higher weights to more informative modalities. In contrast, the early fusion model treats all modalities equally, which may result in noise or irrelevant features from some modalities interfering with the prediction accuracy.

Another advantage of the weighted late fusion model is its ability to capture complex relationships and interactions between different modalities. For example, in emotion recognition, facial expressions in the image modality

may be more informative for detecting happiness, while tone of voice in the audio modality may be more informative for detecting anger. The weighted late fusion model can integrate this information in a more nuanced way than the early fusion model, leading to better performance.
Finally, the weighted late fusion model can also reduce the impact of noisy or unreliable information from individual modalities, by down-weighting their contribution to the final prediction. This important concept was connected to real-world applications, where the quality of data from different modalities might vary significantly.

Finally, the weighted late fusion model can also reduce the impact of noisy or unreliable information from individual modalities, by down-weighting their contribution to the final prediction. This important concept was connected to real-world applications, where the quality of data from different modalities might vary significantly.

iii. *Experiment 3:* The third experiment compared the performance of DHWLF proposed method with other state-of-the-art methods in multimodal emotion recognition. Based on the result, it shows that the proposed method performed the best among other state-of-the-art.

The success of the proposed method can be attributed to several factors. Firstly, the proposed method employs a hierarchical architecture that combines the features from multiple modalities in a weighted manner. This approach allows the model to effectively integrate and utilize the information from each modality for emotion recognition.

Secondly, the proposed method utilizes a deep learning framework which allows the model to extract the most important features from the modalities and fuse them effectively to achieve better results. The approach helps to capture complex patterns in the data, enabling it to recognize emotions more accurately.

Thirdly, the superior performance of the proposed method lies in its fusion strategy. The weighting factor used is based on the performance of each

individual modality in recognizing emotions. This fusion strategy allows the model to capture complementary information from each modality and give more importance to the modalities that are better at recognizing emotions while reducing the influence of the modalities that are less accurate.

In contrast, some of the existing state-of-the-art methods used simple fusion strategies, such as concatenation or averaging of features from different modalities. These methods did not consider the significance of each modality in recognizing emotions, which could lead to the inclusion of irrelevant or redundant features, resulting in reduced accuracy

4.4 Summary of Chapter 4

This chapter presents a detailed design of the DHWLF model for emotion recognition, which integrates three modalities: audio, text, and image. The DHWLF model employs a weighted late fusion approach to combine these modalities with weights determined based on the performance of individual models. The chapter outlines the methodology used to calculate the weights for each modality and explains how this weighting strategy enhances the model's ability to prioritize more reliable signals, resulting in more accurate and robust predictions.

In addition, this chapter provides an overview of the individual emotion recognition models designed for each modality. The SMOTE-2DCNN is for audio modality, TextBERT for the text modality, and Fer2DCNN is for video image modality. These models were trained independently before their outputs were fused using the DHWLF approach.

Three experiments were designed and conducted to evaluate the performance of the DHWLF model. The results and analysis of these experiments are also presented in this chapter, demonstrating the effectiveness of the DHWLF model in outperforming individual models, the early fusion approach and other state-of-the-art methods.

In summary, the success of the proposed method is attributed to the effective integration of the factors discussed in the analysis and discussion section. This approach facilitates the efficient fusion of multimodal data, enabling the extraction of highly discriminative features, which leads to a significant improvement in the overall performance of emotion recognition.

CHAPTER 5

SMOTE-2DCNN FOR AUDIO EMOTION RECOGNITION MODEL

5.1 Introduction

In this chapter, the focus was given to audio modality of data. The main objective of this chapter is to develop an audio emotion recognition model to recognize emotion from humans' speech. There is no doubt that the speech signal contains a significant amount of hidden information which reflects emotion characteristics. The procedure begins with the pre-processing of IEMOCAP audio data, designing of audio emotion recognition model followed by training and testing of the model.

The main issue that will be addressed in this chapter is the imbalance class between four selected labels chosen for this study from the IEMOCAP dataset; (angry, happy, sad and neutral). As mentioned in Chapter 3, there are severe imbalance class between the majority angry class and minority happy class. This issue has become crucial as the imbalance class would dramatically skew the performance of the classifier thus leading to the difficulty of achieving good results in classification. This consequence happened as it could exhibit bias towards the majority class and ignore the minority class altogether. As a result, the model will be overfit and the accurate result might not be achieved for the audio classification.

To overcome this issue, this study includes the use of the SMOTE method in the audio emotion recognition model. SMOTE stands for Synthetic Minority Oversampling Technique which has the ability to solve the problems that occur when using an imbalanced dataset. Further description will be presented in section 5.3 of this chapter.

5.2 Audio Data Pre-Processing

This section provides a detailed explanation of the audio data pre-processing stage, which is essential before training and testing the model. A key challenge in this process is identifying the most relevant audio features for effective emotion recognition. Since this study adopts a deep learning approach for developing the emotion recognition model, it utilizes Mel-Frequency Cepstral Coefficients (MFCC) as the chosen audio representation. MFCC features are well-suited for this task as they allow the model to automatically extract patterns, eliminating the need for manual feature engineering before feeding the data into the deep learning model.

Apart from that, MFCC is a well-known feature that is extracted from audio speech signals that are employed in emotion recognition tasks. The human ear is known to function as a filter, focusing primarily on particular frequency components. In a similar manner, MFCC are understood to represent the filter (vocal tract) and is able to compress information about the vocal tract (smoothed spectrum) into a small number of coefficients based on an understanding of the human cochlea, an ear organ that vibrates at various locations depending on the frequency of external sounds.

Additionally, the coefficient number of MFCC contains information about the rate changes in different spectral bands. If the MFCCs have a positive value, then most spectral energy is present at low frequencies and if the MFCCs have a negative value, most spectral energies are concentrated at high frequencies.

5.2.1 Load Audio Signal Data

The audio wav files generated by this dataset have a sample rate of 22 KHz. This study mostly used Librosa library to process the audio wav files as it is one of the greatest Python libraries in audio processing. The first step of the audio data pre-processing is to load the audio wav files with Librosa and visualize the sound wave to get a clearer overview of the audio data. To build the audio signal data, the sound wave needs to be sampled at the regular time interval and the intensity of the wave for each sample needs to be quantified. The sample rate of the audio signal was standardized once again to the same sampling rate of 44100Hz in order to obtain the same dimension for all arrays. This conversion ensured compatibility with the feature extraction pipeline, which operates optimally at this sample rate to compute features such as Mel-spectrograms and MFCCs accurately (Li et al., 2020). Increasing the sample rate from 22 kHz to 44.1 kHz does not expand the bandwidth of the original signal, as the Nyquist frequency of the 22 kHz signal is limited to 11 kHz. According to the notion, human ear can hear between 20Hz to 20000Hz and therefore, the selection of a standard sample rate that is more than double of human hearing range (44100Hz) is expected to be able to produce a very accurate reproduction (Vomberg, 2019). All the samples were then resized by truncating them to be the same length. Figure 5-1 visualizes the audio time series of IEMOCAP dataset.

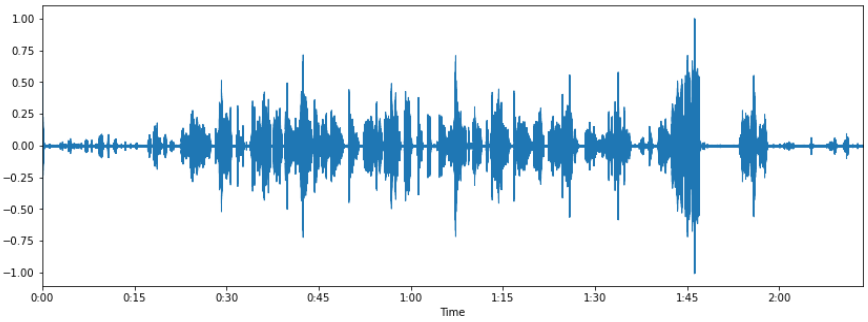

Figure 5.1 : Visualization for Audio Time Series of IEMOCAP Dataset

5.2.2 MFCC Feature Extraction

In theory, to obtain MFCC, it is necessary to go through the processes described below. The first step is to sample the input with frames of size n_ftt=2048 and hop_length=512. From each frame, the frequency spectrum which is also called Short-Time Fourier-Transform (STFT) was calculated in order to convert from time domain to frequency domain. The power spectrum of each frame was then computed using a periodogram, which was modelled after the human cochlea.

Formula to calculate Discrete Fourier Transform (DFT):

$$S_i(k) = \sum_{n=1}^{N} s_i(n)h(n)e^{-j^{2\pi}kn/N}$$

(Equation 15)

where, $S_i(k)$ is the framed time signal, N is the number of samples in a Hamming Window, $h(n)$ is the Hamming Window and k is the length of DFT.

Formula to calculate the periodogram estimate of the power spectrum:

$$Pi(k) = \frac{1}{N}|s_i(k)|^2$$

(Equation 16)

Next, the mel spaced filterbank will be applied to the power spectra. MEL-spaced filterbank has filter spacing that increases exponentially as frequency increases. Each filterbank will be multiplied by the power spectrum to determine the filterbank energies, and the coefficients will then be added. When this is done, it will produce numbers showing how much energy was in each filterbank. The logarithm of the entire series of energy contained in those filterbanks will be calculated in order to get the log filterbank energies. Finally, those numbers of log filterbank energies should then be applied to a Discrete Cosine Transform in order to decorrelate the energies of the overlapping filterbanks. As a result, a set of coefficients were generated which is also known as MFCC.

Despite all the theoretical procedures mentioned above, MFCC can be easily extracted using the Librosa library. Starting with importing all required libraries including Librosa, Pandas, Numpy, Matplotlib, etc. It is then followed by importing a .csv file which contains the name of the audio files and their corresponding classes. The "Iterrows()" function was used to iterate through the pandas dataframe's rows one at a time and run operations on each row sequentially. The "feature.mfcc()" function of librosa with the argument of audio data and appropriate sample rate of the audio signal was simply run in order to finally obtain the MFCC features. In this study, 128 mfccs were extracted; n_mfcc=128 to capture a more detailed representation of the audio signal, considering the nature of the task and the model's requirements. This choice

aligns wit the study by (Ancilin & Milton, 2021), who demonstrated that increasing the number of MFCC coefficients can enhance feature representation for tasks beyond speech recognition. While typical settings for n_mfcc are around 20 (primarily used for speech recognition), emotion recognition requires capturing subtler variations in the audio signal. Increasing n_mfcc provides a finer resolution of the frequency domain, enabling the model to learn more nuanced features related to emotional cues in the audio. After printing the MFCC, the output will seem like an array of numbers, representing the MFCCs for each individual audio sample.

5.2.3 Delta and Delta-Delta Feature Extraction

Speech carries valuable information in its dynamic changes over time, rather than just in individual frames. While MFCC feature vectors capture the power spectral envelope of a single frame, they do not account for temporal variations. To address this, delta (differential) and delta-delta (acceleration) coefficients are introduced, as they track how MFCC values evolve over time. Incorporating these coefficients enhances speech recognition by providing insights into the temporal dynamics of speech. It demonstrates that appending MFCC trajectories to the original feature vector significantly improves recognition performance. The following formula is used to compute the delta coefficients:

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2}$$

(Equation 17)

where, d_t is the delta coefficient, t is the frame considered and C_{t-n} is the coefficients time

Note that the same formula is used to calculate the Delta-Delta coefficients, but this time they are derived from the deltas rather than the static coefficients.

Similar to MFCC, delta and delta-delta coefficients were also calculated using Librosa library by simply running the "feature.delta()" function with order 1 and order 2 for delta and delta-delta features, respectively. Figure 5-2 shows the visualization of MFCCs with delta and delta-delta features derived from this study.

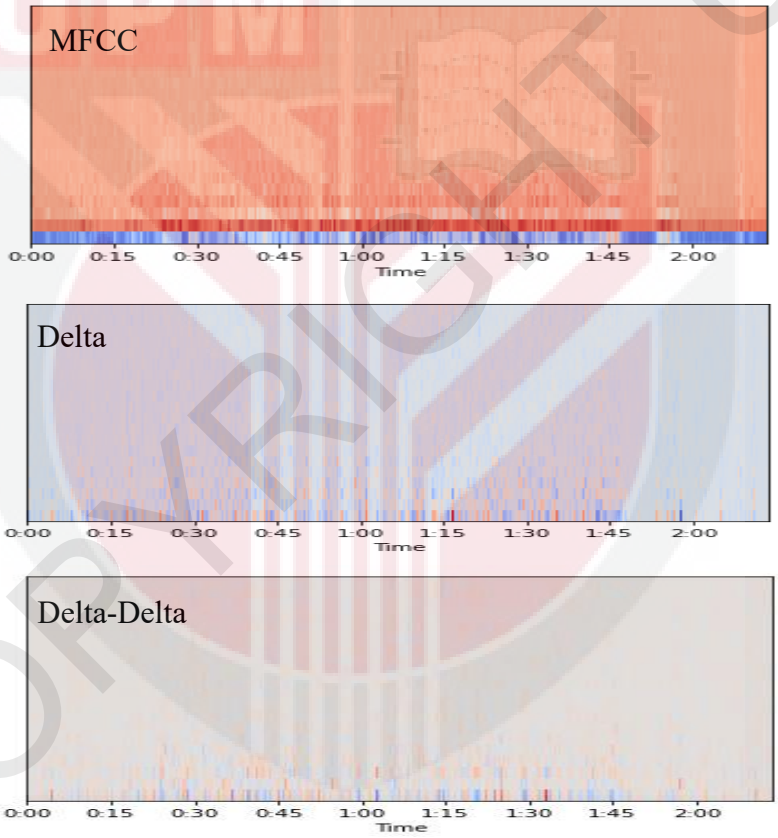

Figure 5.2 : Visualization of MFCCs with Deltas and Deltas-Deltas Features

5.3 Proposed Model of SMOTE-2DCNN

This section describes the proposed audio emotion recognition model. The framework of the audio emotion recognition model is straightforward and simple in design yet

highly functional. This study introduced SMOTE-2DCNN whereby a 2-dimensional convolutional neural network (2DCNN) is implemented with a novel oversampling algorithm called Synthetic Minority Oversampling Technique (SMOTE). The imbalance class still exists even though the up-sampling approaches were employed at the beginning of this study. Therefore, the SMOTE-2DCNN was designed with the purpose to improve the classification and overcome the learning bias brought by an imbalanced class of the dataset.

5.3.1 SMOTE-2DCNN Detailed Description

The SMOTE-2DCNN model consists of two primary modules which are imbalance handling and classification. The imbalance handling module focuses on resampling the training dataset to minimize bias caused by class imbalance in the original data to ensure more reliable experimental results. To achieve this, the model incorporates the SMOTE technique for effective data balancing. Meanwhile, the classification module integrates 2DCNN to enhance the model's performance. The overall structure of the proposed SMOTE-2DCNN model is depicted in Figure 5-3.

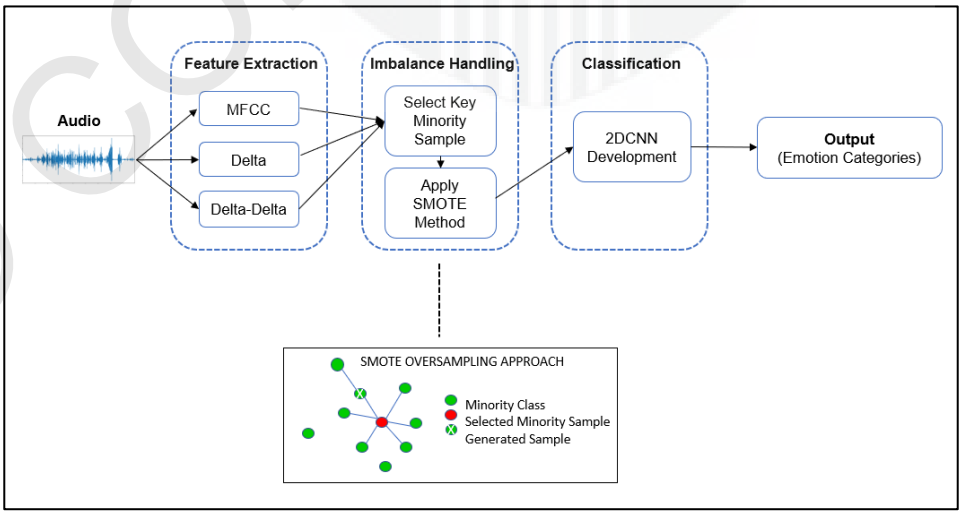

Figure 5.3 : Illustration of Proposed SMOTE-2DCNN

A. Imbalance Handling Module with SMOTE

SMOTE was first introduced in 2002 by Nitesh Chawla (Chawla et al., 2002b). SMOTE is an oversampling technique that creates new synthetic data for the minority class in order to balance the distribution of classes in a dataset. The typical oversampling method has the drawback of not providing the learning model with any additional information or variance, despite the fact that the volume of samples grows. As opposed to that, rather than simply producing duplicates, SMOTE generates synthetic data points that slightly deviate from the original data points.

In order to generate new synthetic data, SMOTE employs the k-nearest neighbour technique. The deep insight of the SMOTE algorithm works in three steps. The first step starts by setting a minority class with set A_{minor}. For each $x \in A_{minor}$, the k-nearest neighbours of x are determined by calculating the Euclidean distance between x and each other element in set A_{minor}. The second step involves the construction of set A_1 by selecting N samples (i.e., x_1, x_2,...x_N) at random from each $x \in A_{minor}$'s k-nearest neighbours. The third step which is the final step, the following formula was used to generate a new synthetic example:

For each $x_k \in A_1$ ($k=1,2,3,\ldots N$),

$$x' = x + rand(0,1) * |x - x_k| \quad \text{(Equation 18)}$$

Each of the aforementioned steps should be repeated until both the minority and majority classes are distributed equally.

B. Classification Module with 2DCNN

A 2D Convolutional Neural Network (2DCNN) is a type of neural network designed to process data structured in two-dimensional arrays, such as images or data with both height and width dimensions. It utilizes a series of filters, also known as kernels or weights, which are typically smaller than the input dimensions, such as 3×3 or 5×5. These filters scan the input data to extract spatial patterns. In this study, the 2DCNN takes a two-dimensional input representing time and extracted features including MFCC, Delta, and Delta-Delta coefficients. The convolution operation applied to these inputs is mathematically expressed by the following equation:

$$y = x * h(n_1, n_2) = \sum_{k1=\infty}^{\infty} \sum_{k2=\infty}^{\infty} x(k_1, k_2) h(n_1 - k_1, n_2 - k_2)$$

(Equation 19)

where, $h\ (k_1, k_2)$ represents the filter and $x(k_1, k_2)$ indicates the input vector of this convolutional layer. The filter $h(k_1, k_2)$ is first turned into $h(-k_1, -k_2)$ and translated by n_1 and n_2 which causes the filter $h(k_1, k_2)$ to eventually become filter $h(n_1 - k_1, n_2 - k_2)$. This is what the negative sign in the filter $h(n_1 - k_1, n_2 - k_2)$ means. Finally, multiply the input $x(k_1, k_2)$ with the result to obtain the value $y(n_1, n_2)$.

After each convolutional layer, an activation function is applied to introduce non-linearity and improve the model's ability to capture complex patterns. The Rectified Linear Unit (ReLU) is chosen as the activation function to mitigate the vanishing gradient problem. The mathematical representation of the ReLU function is provided in the following equation:

$$f(x) = \{x, 0, \quad x > 0 \ \ or\ otherwise$$

(Equation 20)

Each convolutional layer concludes with a max-pooling layer which extracts significant features by selecting the maximum value within a predefined region. It is essential in order to reduce redundancy in the input vector. In this study, a 2 × 2 max-pooling filter was utilized with a default stride length of 2 to match the filter size. In addition, batch normalization was incorporated between layers to standardize inputs, enhancing training stability and accelerating convergence. The batch normalization process is mathematically represented by the following equations:

$$\mu_B \leftarrow \frac{1}{3}\sum_{i=1}^{m} x_i \quad \text{(Equation 21)}$$

$$\sigma_B^2 \leftarrow \frac{1}{m}\sum_{m}^{i=1}(x_i - \mu_B)^2 \quad \text{(Equation 22)}$$

where the mean and variance of batch data input are represented by μ_B and σ^2_B. Each dimension of input is then separately normalized with the implementation of following formula equation:

For $x = (x_1 \ldots . x_d)$,

$$\hat{x}_i \leftarrow \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}} \quad \text{(Equation 23)}$$

For numerical stability, a tiny constant ε is added to the denominator in order to avoid the occurrence of dividing by value 0. By scaling and shifting the regularized value data $\hat{X}_i$, it allows the batch normalization to reinstate the power of the network. This process can be expressed by the following formula equation:

$$y_i \leftarrow \gamma \hat{x}_i + \beta$$

(Equation 24)

where ɣ and **β** are the parameters that later will be learned during the optimization process.

As the network gets deeper, a flatten layer is applied to combine all layers into one single layer, transform each dimensional array into a single lengthy continuous linear vector and provide them as inputs to the subsequent fully connected layers, also known as dense layers. A couple of fully connected layers have been added to wrap up the classification model. By multiplying the input vector by the weights matrix and then adding the bias vector, the fully connected layers are in charge of creating a linear transformation to the input vector. The product is then subjected to a non-linear transformation using the non-linear activation function *f* which is ReLU. The process is repeated for each fully connected layer and the calculation of this process can be expressed by the following equation:

$$y = f(Wx + b)$$

(Equation 25)

where, *x* is input vector, *W* is weight vector, *b* is bias and *f* is activation function.

To prevent overfitting, a dropout layer is incorporated as a regularization technique. After the fully connected layers, the final layer utilizes the softmax activation function to compute the probability distribution over the possible classes. The equation for the softmax activation function is presented below:

$$softmax\ (z)_i\ = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}}$$

(Equation 26)

Apart from that, the categorical cross-entropy was employed as loss function for training model in which can be expressed by the following equation:

$$Loss = \frac{-1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} y_{i,j} \log_2(\hat{y}_{i,j})$$

(Equation 27)

where, N denotes number of samples, C is the number of classes. y_i is the one-hot encoded vector representing the true class label of the input sample in the form of 0s 1s and $\hat{y}_{i,j}$ is a vector representing the predicted probabilities for each class.

5.3.2 2DCNN Model Architecture

This section provides a detailed explanation of the 2D CNN model architecture. With a hop length of 512 (page 140), an audio duration of approximately 2.25 seconds (Figure 5-2) and typical settings for MFCC extraction (e.g., 13 coefficients), the initial spectrogram results in a matrix size of approximately 128 × 190. To adapt this input for the CNN and reduce computational complexity, the MFCC features including delta and delta-delta features were down-sampled and reshaped. The resulting representation was reshaped into a fixed size of 16 × 8 × 1, which is compatible with the CNN's input layer. This smaller input size was selected to meet the design requirements of the CNN, hence ensuring a fixed input size for both training and inference. The architecture of the 2D CNN network is outlined below:

- **CONV 1:** The initial convolutional layer processes an input of dimensions 16 × 8 × 1. It employs 64 filters, each with a spatial size of 3 × 3. A Leaky Rectified Linear Unit (ReLU) serves as the activation function, and the layer utilizes "same" padding. This is followed by a max-pooling operation with a 2 × 2 stride and the application of Batch Normalization.

- **CONV 2:** The second layer consists of 64 filters, each with a spatial size of 3 × 3. It employs the Leaky Rectified Linear Unit (ReLU) activation function and uses "same" padding. This layer is followed by a max-pooling operation with a 2 × 2 stride and Batch Normalization

- **CONV 3:** The third layer utilizes 64 kernels, each with a spatial size of 3 × 3. It employs the Leaky Rectified Linear Unit (ReLU) activation function with “same” padding. This is followed by a max-pooling operation with a 2 × 2 stride step.

- **DENSE 1:** The first dense layer consists of 64 units and utilizes the Leaky Rectified Linear Unit (ReLU) as its activation function. It incorporates both kernel and bias regularizers, applying L2 regularization with a default value of l2=0.01. A dropout rate of 0.5 is included in this layer.

- **DENSE 2:** The final dense layer consists of four units and utilizes the Softmax activation function.

Figures 5-4, 5-5, and 5-6 illustrate the proposed 2DCNN model architecture, the model summary, and the graphical representation of the proposed 2DCNN model, respectively.

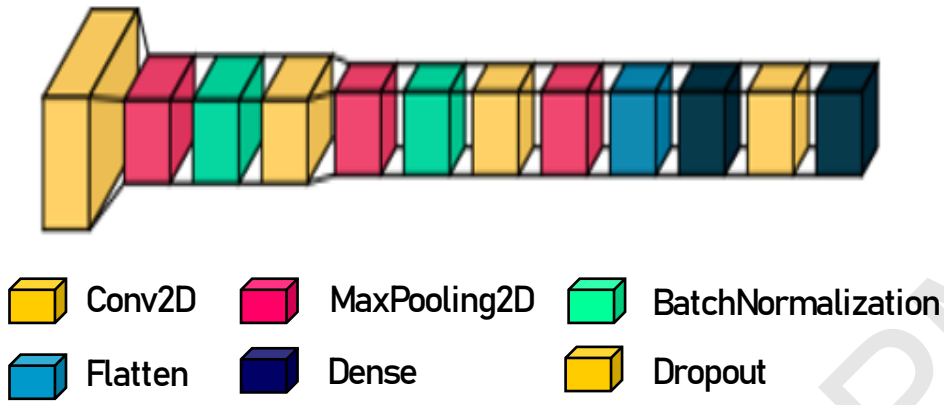

Figure 5.4 : Proposed SMOTE-2DCNN Model Architectures

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 16, 8, 64)	640
max_pooling2d (MaxPooling2D)	(None, 8, 4, 64)	0
batch_normalization (BatchNormalization)	(None, 8, 4, 64)	256
conv2d_1 (Conv2D)	(None, 8, 4, 128)	73856
max_pooling2d_1 (MaxPooling2D)	(None, 4, 2, 128)	0
batch_normalization_1 (BatchNormalization)	(None, 4, 2, 128)	512
conv2d_2 (Conv2D)	(None, 4, 2, 64)	73792
max_pooling2d_2 (MaxPooling2D)	(None, 2, 1, 64)	0
Flatten (Flatten)	(None, 128)	0
dense (Dense)	(None, 64)	8256
dropout (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 4)	260

Total params: 157,572
Trainable params: 157,188
Non-trainable params: 384

Figure 5.5 : Proposed SMOTE-2DCNN Model Summary

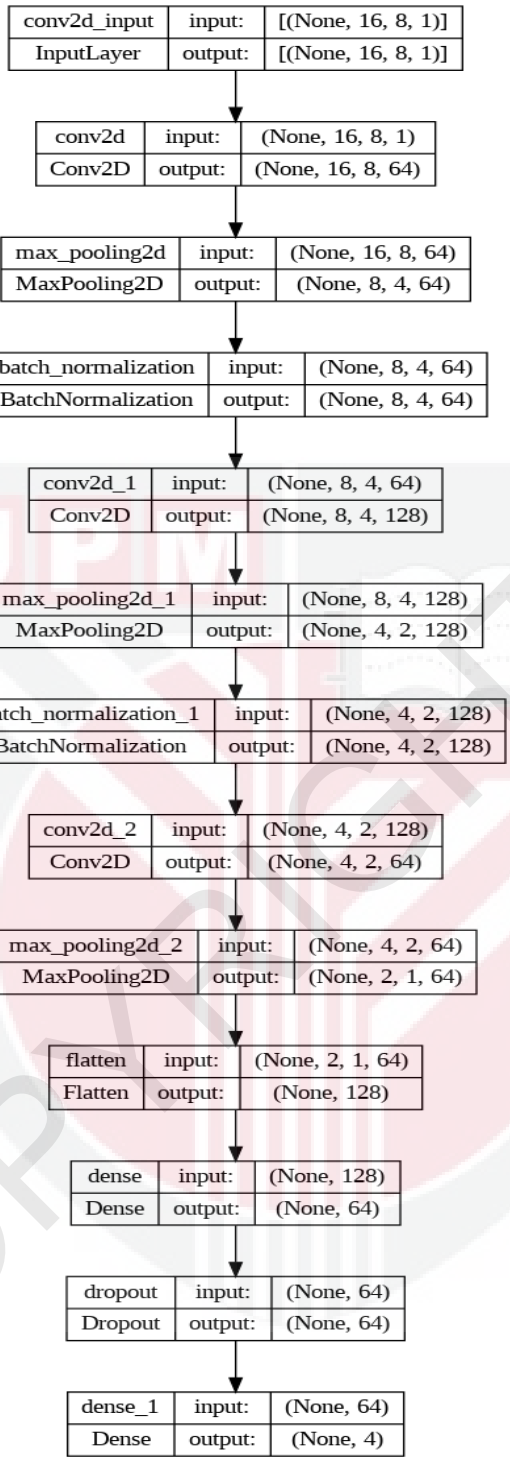

Figure 5.6 : Plot of Proposed SMOTE-2DCNN Model Graph

5.3.3 SMOTE-2DCNN Model Algorithm

The proposed SMOTE-2DCNN model's steps are outlined in Algorithm 2. Let D be the dataset, where $D = \{(d_1, y_1), \ldots, (d_A, y_A)\}$. Here, (d_A, y_A), represents the feature set and label of audio data A. For $x \in A_{minor}$, the k-nearest neighbors of x were determined and synthetic samples A_1 were generated using Equation 18 to balance the class distribution. To avoid data leakage, the dataset was split into training, validation, and testing sets with an 80:10:10 ratio. Synthetic samples were generated exclusively from the training set, ensuring that neither the validation nor the testing sets were contaminated with artificially created samples. The validation set was used for hyperparameter tuning and performance monitoring during training, while the testing set was reserved for final evaluation.

The next step involves constructing the 2DCNN model by defining its architecture. This includes adding convolutional layers, max-pooling layers, flattening the output from the convolutional layers, incorporating dense layers and applying the softmax activation function for classification. The model is then compiled with an appropriate loss function and optimizer before being trained using backpropagation and gradient descent on a labeled training set. The details of hyperparameter optimization are provided in Section 5.4.1, Table 5-2. Finally, the model's performance is evaluated on the testing set, with the predicted class serving as the final output.

Algorithm 2: SMOTE-2DCNN Model	
Input:	$D = \{(d_1, y_1), \ldots, (d_A, y_A)\}$ A_{minor} = Minority class instances, k = Number of nearest neighbors to use, N = Number of synthetic samples to generate
Output:	A_1 =Synthetic minority class samples $FinalOutput_{Test}$ = Label class predicted on the provided testing sample, *Test* acc_{M1} = Percentage accuracy of M_1 fusion model M_1 represent model 1 which is the proposed SMOTE_2DCNN.
1) Split D into training set (D_{train}), validation set (D_{val}) and testing set (D_{test})	
2) For each sample x in A_{minor} from D_{train} ;	Compute the k- nearest neighbors of x within A_{minor}, excluding x itself.
3) For k = 1 to N :	a) Choose a random neighbor x_k from the k nearest neighbors of x. b) Generate a new synthetic sample x' by following Equation 18. c) Add the new synthetic sample x' sample to Synthetic minority class samples, A_1 d) Return the updated training set: $D_{train} = D_{train} \cup A_1$
4) Define 2DCNN Model architecture	
5) Initialize weights and biases randomly	
6) Set hyperparameters; learning_rate, num_epochs, batch_size, Optimizer	
7) Train model: For epoch= 1 to num_epochs	a) Loop over batches of training data, D_{train} b) Extract a batch of input data and corresponding labels c) Perform forward pass using Equations 19-27 d) Compute the loss e) End For
8) Perform backward pass: For layer in reversed layers (starting from last layer moving backwards)	a) Compute gradients for each layer b) Update layer weights and bias using gradients and optimizer c) Validate the model on D_{val} after each epoch and monitor metrics e) End For

9) Evaluate the trained SMOTE-2DCNN model on D_{test} to compute the predicted class labels, $FinalOutput_{Test}$, where Ti (i = 1…nc with nc=4, the number of classes*)* represents the predicted label for the i^{th} test sample.
10) Calculate the metrics such as accuracy, precision, recall, and F1 score
11) Compute the final percentage accuracy of proposed SMOTE 2DCNN model, acc_{M1}
12) End

Algorithm 2 is distinct from methods described in the anchor papers in its approach to synthetic data generation, dataset handling, and feature transformation. Unlike the methods of (Tripathi et al., 2019) , (J. Zhang et al., 2022),(H. Xu et al., 2019) and (Santoso et al., 2022), this algorithm integrates a modified SMOTE technique tailored to audio emotion recognition tasks. For instance:

- Synthetic Data Generation: Algorithm 2 generates synthetic samples using *k*-nearest neighbours and Equation 18, specifically for the training set. The anchor papers primarily process raw audio features without explicitly addressing class imbalance through synthetic data.
- Feature Transformation: While Algorithm 2 down-samples and reshapes MFCC features into 16×8×1, anchor papers retain full-dimensional raw features.

Example Calculation for the SMOTE of Algorithm 2:

Consider a dataset D containing audio samples with a minority class A_{minor} :

1. A sample $x \in A_{minor}$ is selected, e.g., x= [1.2,0.8,1.5,2.0]
2. Compute k-nearest neighbours (k=3): x_1= [1.1,0.9,1.4,2.1], x_2= [1.3,0.7,1.6,2.2], x_3= [1.0,0.8,1.5,2.0]
3. Generate a synthetic sample using Equation 18 with $rand$=0.5 and $x_k = x_1$:

$x' = x + rand\ (x_k - x)$=[1.15,0.85,1.45,2.05]

4. Add x' to D_{train}

This process is repeated for all samples in A_{minor} to balance the training set.

5.4 Evaluations and Results

This section provides a detailed discussion of the evaluation procedure and results of the SMOTE-2DCNN model. The proposed model was assessed using the IEMOCAP audio dataset. To ensure consistency with prior studies, this research focused on four specific emotion categories namely angry, happy, sad and neutral which have been selected from the annotated categorical labels in the IEMOCAP dataset.

The dataset was split into training, validation, and testing sets using the "train_test_split()" function, following an 80:20 ratio. In this setup, 80% of the data was allocated for training, while the remaining 20% was used for validation and testing. To enhance model performance, stratified k-fold cross-validation was employed by enabling the stratify parameter, ensuring a balanced distribution of target classes across all folds. This approach helps minimize bias and improve classification accuracy. Table 5-1 presents the dataset division details after splitting.

Table 5.1 : Details of Dataset Division (Audio)

IEMOCAP (Audio) *(Involving 4 label class; angry, happy, sad, and neutral)*	
Number of Samples for Training Set	5465
Number of Samples for Validation and Testing Set	1367

5.4.1 Hyperparameter Optimization for SMOTE-2DCNN

Before conducting any experiments, it is essential to identify the optimal hyperparameters for the proposed SMOTE-2DCNN model. Hyperparameter tuning plays a critical role in the model's performance, as it can significantly affect its accuracy. The objective of hyperparameter tuning is to determine the best set of values that will maximize the model's accuracy on the test set.

Several key hyperparameters were considered during the tuning process for the proposed SMOTE-2DCNN. Table 5-2 provides a summary of the best-tuned hyperparameter values for the model.

Table 5.2 : Hyperparameters Tuning Value (SMOTE-2DCNN)

Used Hyperparameter	Values
Number of Epochs	100
Batch Size	64
Learning Rate	0.01
Kernel	3
Dropout Rate	0.5
Optimizer	Adam
Regularized	kernel regularizer = L2 (0.01) bias regularizer = L2 (0.01)

5.4.2 Experimental Set Up for SMOTE-2DCNN

This section presents a comprehensive evaluation comprising three experimental designs to assess the effectiveness of the proposed SMOTE-2DCNN model.

The first experiment focuses on training the 2DCNN model using two distinct feature input sets. In the first set, MFCC, Delta, and Delta-Delta features are computed, and their mean values are taken to reduce feature dimensionality. In contrast, the second

set includes the full arrays of MFCC, Delta, and Delta-Delta features without averaging.

The second experiment examines the impact of the SMOTE approach by comparing the performance of the 2DCNN model with and without SMOTE. The model is trained using the best feature set from the first experiment and the results are analyzed for differences in classification performance.

The third experiment evaluates the proposed SMOTE-2DCNN method against state-of-the-art models in audio emotion recognition. To ensure a fair comparison, all competing methods are trained using the same IEMOCAP dataset.

5.4.3 Result

This section presents the evaluation results and findings of the proposed SMOTE-2DCNN on the aforementioned sets experimental design.

A. Experiment 1: Comparison Between Two Feature Input Sets

The first experiment was conducted to verify the best form of input set feature. By comparing different forms of input feature sets will contribute more performance improvement to the proposed model.

Table 5.3 : Comparison of Different Form of Input Feature Set

Input Feature Set	Accuracy (%)
Mean of MFCC, Delta, Delta-Delta	70
Entire Arrays of MFCC, Delta, Delta-Delta	75

Table 5-3 presents the experimental results comparing various input feature sets. The 2DCNN model achieves 75% accuracy when utilizing the entire arrays of MFCC, Delta, and Delta-Delta, surpassing the 70% accuracy obtained when using only the mean values of these features.

B. Experiment 2: Comparison Between with SMOTE and without SMOTE

The set of MFCC, Delta, and Delta-Delta feature with entire arrays was selected for the final experiment with the SMOTE-2DCNN model, as it yielded the highest accuracy in the previous evaluation.

To demonstrate the effectiveness of the proposed approach, a performance comparison was conducted between the 2DCNN model with and without SMOTE. Table 5-4 presents the accuracy percentages for both versions of the model. In addition, Figure 5-7 visualizes the accuracy and loss per epoch for the 2DCNN model with SMOTE, while Figure 5-8 provides the same representation for the model without SMOTE.

Table 5.4 : Percentage Accuracy of 2DCNN with and without SMOTE

Method	Accuracy (%)
With SMOTE	80
Without SMOTE	75

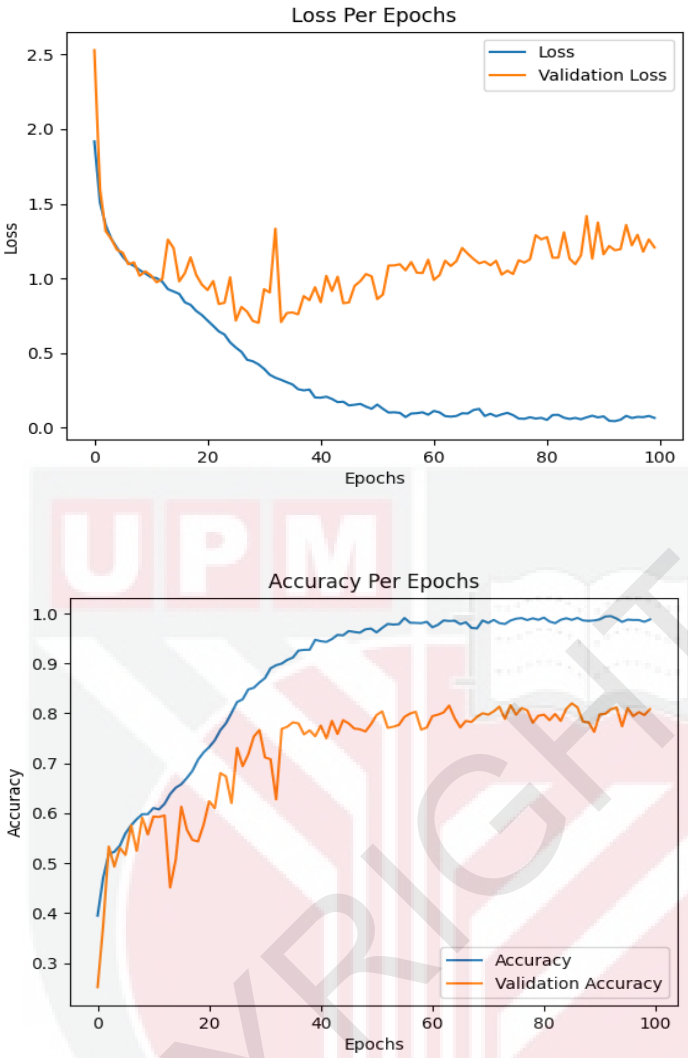

Figure 5.7 : Accuracy Per Epoch and Loss Per Epoch of 2DCNN with SMOTE

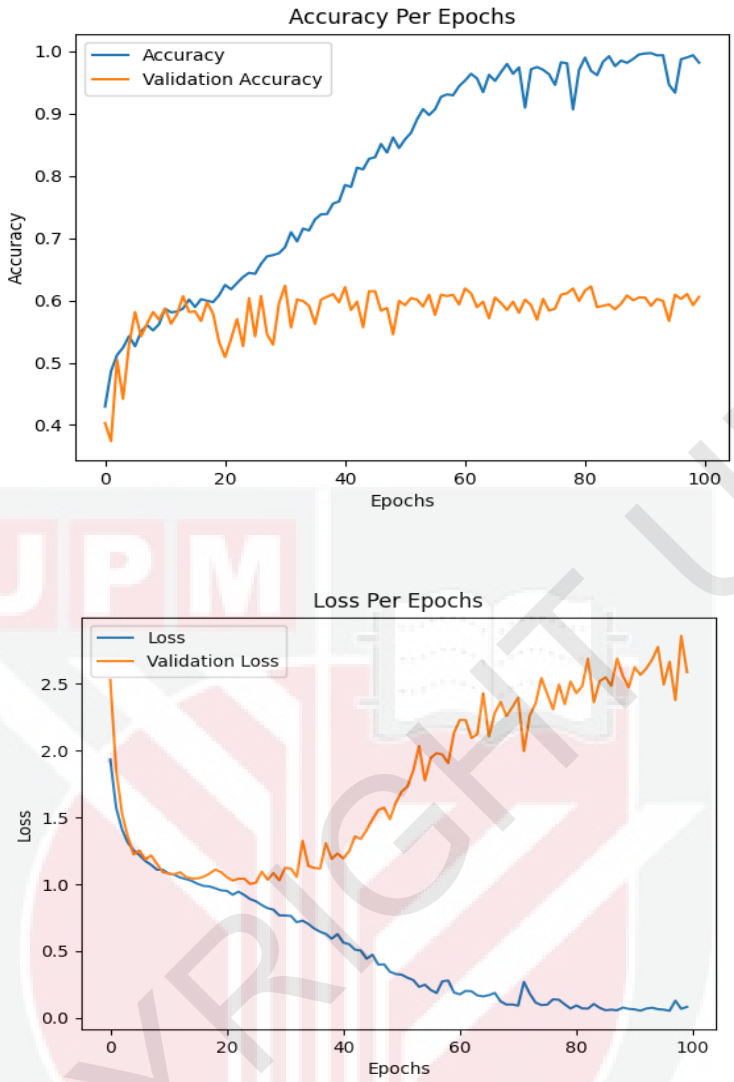

Figure 5.8 : Accuracy Per Epoch and Loss Per Epoch of 2DCNN without SMOTE

As shown in Figure 5-9, the proposed SMOTE-2DCNN achieves an accuracy of 80% with a corresponding loss of 1.3. This represents a notable 5% improvement over the 2DCNN model without the SMOTE method. Additionally, the confusion matrix in Figure 4-9 indicates that the model demonstrates a high True Positive Rate (TPR) for most emotions, achieving 82% for angry, 90% for happy, and 85% for sad. However, its performance is slightly lower for the neutral emotion, with a TPR of 71%.

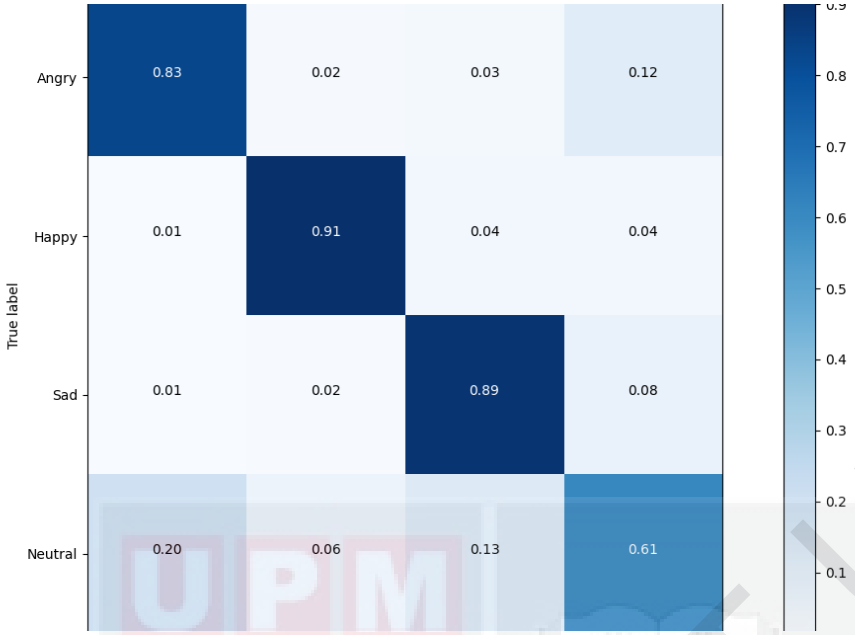

Figure 5.9 : Confusion Matrix of SMOTE-2DCNN

Table 5-5 shows the result of precision, recall and F1 score of each emotion label; angry, happiness, sadness and neutral. Among all, happy emotion scores highest precision with 89%. Same goes for recall and F1-scores, happy emotion scores the highest with 91% and 90%, respectively. This study observes that these results are attributed to the unique acoustic features of happy emotion such as higher pitch, greater energy, and faster speech rate. These attributes create a distinct pattern in the audio features especially in MFCCs and their derivatives (delta and delta-delta coefficients), making them easier for the model to classify accurately.

Table 5.5 : Precision, Recall and F1-Score of SMOTE-2DCNN

Emotion	Precision (%)	Recall (%)	F1-Score (%)
Angry	78	83	81
Happy	**89**	**91**	**90**
Sad	82	89	85
Neutral	72	61	66

C. Experiment 3: Comparison with State-of-The-Art

Table 5-6 and Figure 5-10 provide a comparative analysis of percentage accuracy, including a graphical representation for both the proposed SMOTE-2DCNN model and state-of-the-art methods in audio emotion recognition. To ensure consistency with previous studies, both unweighted accuracy (UA) and weighted accuracy (WA) were used to record the accuracy results. As shown in Table 5-6, the SMOTE-2DCNN model achieves the highest UA and WA scores of 81% and 80%, respectively. This represents a significant improvement of approximately 15% over the best-performing method among the state-of-the-art approaches.

Table 5.6 : Comparison SMOTE-2DCNN with State-of-The-Art

Method	UA (%)	WA (%)
BiLSTM +Attn (Tripathi et al., 2019)	51.2	55.6
BLSTM + SelfAttn (Santoso et al., 2022)	76.8	76.6
LSTM + Attn (Xu et al., 2019)	57.4	63.4
CNN + LSTM (Satt et al., 2017)	59.4	68.0
2TransformerEncoder (J. Zhang et al., 2022)	57.1	63.6
SMOTE-2DCNN	81.0	80.0

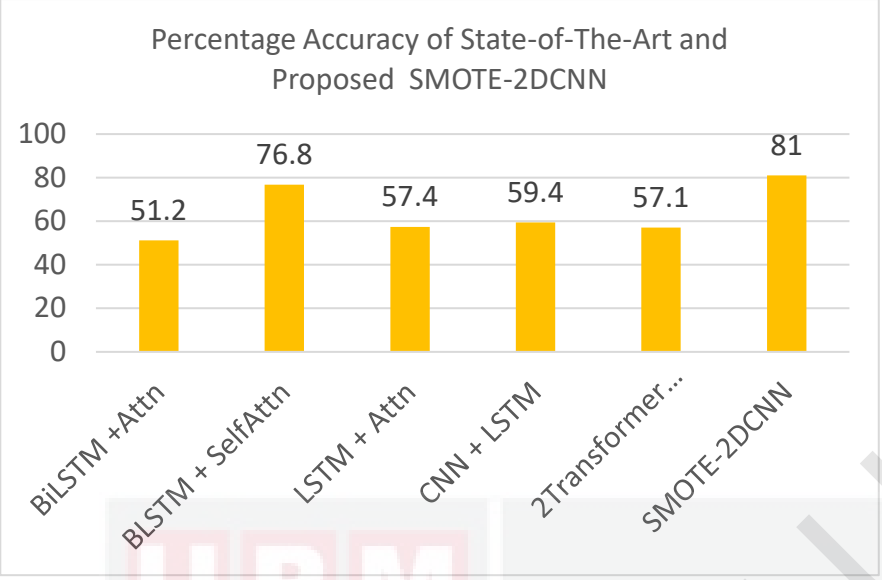

Figure 5.10 : Percentage Accuracy of State-of-The-Art and Proposed SMOTE-2DCNN

5.5 Analysis and Discussion

The proposed SMOTE-2DCNN demonstrates a substantial performance improvement over conventional 2DCNN models and multiple state-of-the-art audio speech emotion recognition approaches (refer Table 5-6). This exceptional performance can be attributed to several key factors which are discussed below:

i. *Simple Model with Effective Imbalance Class Handling Module:* The SMOTE-2DCNN incorporates the SMOTE method to address class imbalance in the IEMOCAP dataset. Unlike conventional random oversampling methods, SMOTE generates synthetic samples based on feature space similarities rather than duplicating existing minority class instances. Furthermore, it utilizes the k-nearest neighbors' algorithm to determine where synthetic samples should be created in order to ensure they closely resemble the actual minority class data. This technique helps to reduce overfitting, which is a common issue in random oversampling. Hence, it leads to improved classification accuracy and a reduced bias toward majority classes.

ii. *Compatibility on the Classifier and Imbalance Class Handler:* The SMOTE-2DCNN model consists of two key components: The Imbalance Handling Module with SMOTE and the Classification Module with 2DCNN. This combination is highly effective when it can address extreme class imbalance while remaining robust to small sample size variations. SMOTE serves as a reliable technique for handling class imbalance, making it well-suited for enhancing the performance of the 2DCNN model, particularly in emotion recognition tasks. The effectiveness of this approach lies not only in its ability to balance the dataset but also in the synergy between SMOTE and the classifier. Their compatibility is crucial for achieving optimal performance. In other words, SMOTE enhances the representation of minority class instances, enabling the 2DCNN model to achieve greater classification accuracy and robustness.

iii. *Selection of Input Features:* It is prominent to select the best form of input set feature to fit in with the model. Incorporating the entire arrays of MFCC, Delta, and Delta-Delta features prevents misrepresentation and the loss of certain important information from the data. It is therefore contributing to the improvement of the proposed SMOTE-2DCNN model performance especially in terms of validation accuracy and validation loss.

5.6 Summary of Chapter 5

This chapter focused on developing and evaluating the proposed SMOTE-2DCNN model for audio emotion recognition using the IEMOCAP dataset. The model aimed to address the challenges of imbalanced class distribution among the four emotion labels: Angry, Happy, Sad, and Neutral. The chapter began with an overview of the audio data preprocessing steps, including the extraction of MFCC, delta, and delta-delta features, which are critical for capturing the nuances of emotional expressions in speech.

To handle the severe class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was incorporated into the model. This technique generated synthetic samples for the minority class based on feature-space similarities, effectively reducing bias and improving classification performance. The SMOTE-generated samples were added exclusively to the training set, ensuring that the validation and test sets remained untouched to prevent data leakage.

The 2DCNN classifier was specifically designed to process 2-dimensional audio features, combining convolutional layers, max-pooling layers, and dense layers to extract and classify emotional cues from the input data. The results demonstrated that the proposed SMOTE-2DCNN model outperformed the baseline 2DCNN model and state-of-the-art approaches in audio emotion recognition. The inclusion of the SMOTE module and the use of the entire arrays of MFCC, delta, and delta-delta features were key contributors to this success.

Performance metrics such as precision, recall, and F1-score highlighted the model's effectiveness, with the "Happy" emotion achieving the highest scores due to its distinctive acoustic features. The evaluation also revealed that the proposed model achieved significant improvements in unweighted and weighted accuracy compared to other state-of-the-art methods.

In conclusion, the SMOTE-2DCNN model effectively addressed the challenges of imbalanced class distribution and demonstrated superior performance in audio emotion recognition tasks. This study underscores the importance of integrating

imbalance handling techniques with robust deep learning architectures, paving the way for future advancements in emotion recognition systems.

CHAPTER 6

TextBERT FOR TEXT EMOTION RECOGNITION MODEL

6.1 Introduction

In this chapter, the focus was given to text modality of data. The main objective of this chapter is to develop a text emotion recognition model from transcript text data that is associated with human speech. It can be difficult to recognize emotion from text data associated with human speech for a number of reasons. One of the difficulties in recognizing emotion from text is that emotions are frequently expressed indirectly, and the same words or phrases can convey a variety of emotions depending on the context. One of the difficulties in recognizing emotion from text is that emotions are frequently expressed indirectly, and the same words or phrases can convey a variety of emotions depending on the context.

In this chapter, focus will be given in learning the context-dependent between words in order to recognize human emotion from speech text data. One of the primary concerns that will be tackled in this chapter when introducing context into word embedding is to understand the relationship between the words and their context. It is quite confusing as it contains ambiguities that need to be interpreted correctly. The complexity of natural language and the vast number of possible contextual aspects also can make it difficult to design and train models that are capable of capturing context nuances.

To address this issue in text emotion recognition, this study introduces TextBERT which implements the BERT language model in the development of the text emotion

recognition model. Nevertheless, due to its deep architecture and large number of parameters, BERT frequently faces computational intensity challenges, which can require substantial computing resources to train and use effectively. Furthermore, domain adaptation can be a significant challenge because BERT's pre-trained models may not perform optimally in specific domains and it requires fine-tuning or further training to achieve good performance for specialized tasks.

Therefore, this study proposes a lightweight BERT model that maintains good performance despite being a faster model. Further discussion regarding the proposed system will be presented in section 6.2 of this chapter.

6.2 Proposed Model of TextBERT

This study uses the BERT language model to develop a text emotion recognition model known as TextBERT. BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based model that has been previously trained on a large corpus of text data (refer Chapter 2). BERT has the ability to interpret word meaning in context by employing attention mechanism, deep learning techniques to comprehend the context of words in a sentence. It processes the text bidirectionally, taking into account the words before and after a given word in a sentence.

6.2.1 Data Preparation with TextBERT

BERT is able to understand and process text in a manner similar to that of humans. BERT does not require any additional pre-processing of the text data because it can handle the majority of text data pre-processing tasks like tokenization, lowercasing, and handling special characters, changes in capitalization, punctuation, and grammar.

In order to use a pre-trained BERT model, the input data need to be converted into an appropriate format to ensure that each sentence can be transmitted to the pre-trained model and the corresponding embedding can be obtained. There are few specific procedures to be employed. The first step of the procedure is tokenization whereby the texts are split into individual tokens or sub words. BERT uses a WordPiece tokenizer, which divides words into sub words based on their frequency of occurrence in the training data. By doing this, it allows to limit the amount of vocabulary and out-of-vocabulary words will be handled correctly.

Once the input text has been tokenized, each token is assigned a unique numerical ID to be used by the BERT model. This encoding process is crucial because it transforms text into a numerical structure that the model can comprehend. Next, the special tokens have been added whereby [CLS] at the first position and [SEP] at the end of the sentence to denote the start and end of a sentence, respectively. The following step is to pad the sequences such that each sequence has the same length due to the variable length of the text data. This was accomplished by adding unique tokens [PAD] to the start or end of the sequences until they were all the same length.

In order to clearly distinguish between real and [PAD] tokens, a technique called masked language modelling [MASK] has been employed. In this method, 15% of the tokens in the input sequence are randomly masked, and the model is then trained to predict the missing tokens based on context, which contributes to a deeper understanding of the language. Finally, using pre-trained BERT embeddings, all the tokens will be converted into tensors, which are multi-dimensional arrays, and fed through the model to produce the final representations.

6.2.2 Bert Detailed Description

This section describes in detail about the detailed description of the BERT language model used for the text emotion recognition model. Figure 6-1 shows the model framework based on the BERT algorithm.

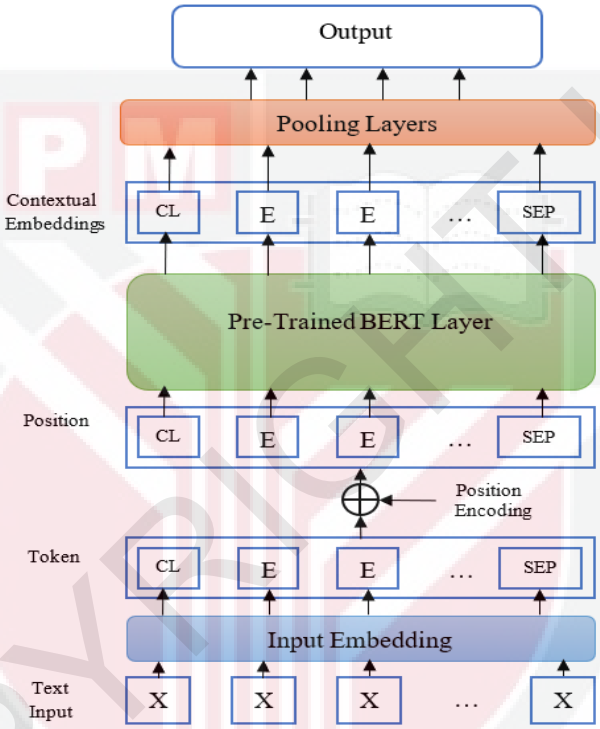

Figure 6.1 : Model Framework Based on the BERT Language Model

A. Sentence Positional Encoding Vector

BERT first converts each word in the text into a token embedding, which is a fixed length vector used for text data processing. However, this is insufficient because the meaning of a word can differs depending on where it appears in a sentence. To capture the contextual information, BERT uses sentence positional encoding vectors to give the model information about the relative positions of words in a sentence. Particularly

in the context of language representation models, this information is essential to understanding the meaning of a sentence.

In order to generate sentence positional encoding vectors in BERT, a vector representation for each word in the sentence must be created. Each vector has a length equal to the dimensionality of the model's hidden layers. The formula to generate the positional encoding vector is shown as follow:

$$PE_{pos,2i} = sin\left(\frac{pos}{1000^{d}\ \frac{2i}{model}}\right)$$

$$PE_{pos,2i+1} = cos\left(\frac{pos}{1000^{d}\ \frac{2i}{model}}\right)$$

(Equation 28)

where, *PE* is the positional encoding vector, *pos* is the position of the word in the sentence, *i* is the dimensionality of the model' hidden layers and d is the number of dimensions of the positional encoding vector.

The formula produces two values for each dimension, one for the sine value and the other for the cosine value. The final sentence positional encoding vector is created by adding these values to the word representation. The sine and cosine functions are used in the formula because it is assumed that the position of words in a sentence follows a cyclical pattern and these functions would be useful in capturing that pattern.

B. Multi-Headed Attention in BERT

BERT (Bidirectional Encoder Representations from Transformers) relies on Multi-Headed Attention to learn context-dependent representations of input sequences. The primary goal of this process is to determine the relationship between each token in a sentence and the relevance of each token in context to the other tokens in the sentence. This enables BERT to capture the contextual relationships between tokens, improving the model's accuracy in language-related tasks like natural language understanding and classification.

The fundamental principle behind multi-headed attention is to divide the attention mechanism into several heads, each of which pays attention to various aspects of the input. The overall process begins when the input sequence is initially integrated into the query (Q), key (K), and value (V) matrices. (More details in Chapter 2). The dot product of the key and query matrices is then used to calculate the attention scores by using the following equation:

$$Attention\ (Q, K, V) = softmax\ (\frac{QK^T}{\sqrt{d_k}})\ V$$

(Equation 29)

In light of the above equation, the three weight matrices ((W(Q), W(K) and W(V)) are multiplied by each encoder input vector. Three vectors; the key vector, the query vector, and the value vector will be produced for each input vector as a result of this matrix multiplication. The query vector of the current input is then multiplied by the key vectors from other inputs and the dot product of both matrices is used to calculate the attention scores. In order to prevent the attention scores from exploding due to

large values in the key matrix, the attention scores are then scaled by a factor of $\sqrt{d_k}$. This is known as scaled dot-product attention. The softmax function is then applied to all self-attention scores calculated in relation to the query word (also known as the first word) before multiplying it with the value vector. All the weighted values obtained in the previous step will be summed up to get the self-attention output of the given word.

During the operation of multi-headed attention, the query, key, and value matrices are linearly projected numerous times, yielding multiple attention scores. Concatenating the output from each head and linearly projecting it to the desired output size yields the multi-head attention system's final output. Multi-headed attention can be represented mathematically by this formula:

$$MultiHead\ (Q, K, V) = concat\ \left(head_{1,}\ head_{2,}\ \ldots., head_n\right) W_o$$

$$where, head_{i,} = Attention\ (QW_i^Q, KW_i^K, VW_i^v)$$

(Equation 30)

To calculate multi headed attention, the embedding of each word of the input sentence needs to be generated first. The weight matrices are differed for each head (W(Q), W(K), and W(V)) and each of them need to be multiplied with the input matrix matrices (W^Q, W^K, and W^V) to produce key, value and query matrices for each attention head. Next, apply the attention mechanism to the query, key, and value matrices whereby it will result in an output matrix for each attention head. Finally, concatenate the output matrices from each attention head and dot product with the weight W_O to create the output of the multi-headed attention layer.

In general, the multi-headed attention allows the BERT model to pay attention to different aspects of the input sequence and capture long-term dependencies between tokens in the input sequence. The model benefits from the multi-headed attention process by better understanding the context and making accurate predictions of masked tokens in the input sequence.

6.2.3 BERT Tensorflow Implementation

This section explains how to incorporate a pre-trained BERT model for text emotion recognition model. The data used is the available text transcription from the IEMOCAP dataset. The text transcription was extracted from each utterance in audio data of IEMOCAP dataset.

The TensorFlow 2.0 library and the BERT pre-trained model were initially installed in order to implement BERT for emotion identification in TensorFlow. Next, the transformers library's BERT model and tokenizer were loaded. The "bert-base-uncased" model from Tensorflow's transformers library was then employed. With the ability to be customized for certain tasks, this model has been pre-trained on a large corpus of text data.

Next, the IEMOCAP dataset was loaded. The dataset consists of written transcripts of audio files with each line of text has a corresponding label for the emotion it represents. The “pandas” library was used to load and pre-process the data. As BERT requires numerical input, the text data was converted into numerical form during the pre-processing stage. The Tokenizer tool from the Keras library of TensorFlow was

utilised to tokenize the text input. The tokenized data was then padded to a fixed length.

In the following step, a fine-tuning model was built by adding a dense layer on top of the BERT model which would use the contextual encoding representations generated by BERT as input to perform the classification task. Following the dense layers, the softmax activation function was employed in order to determine the probability that the input belongs to a particular class. Apart from that, the categorical cross-entropy was employed as the loss function for training the model. Figure 6-2 and Figure 6-3 presents the model summary and plot of other BERT model graphs, respectively.

Layer (type)	Output Shape	Param #	Connected to
input_word_ids (InputLayer)	[(None, 250)]	0	[]
input_mask (InputLayer)	[(None, 250)]	0	[]
segment_ids (InputLayer)	[(None, 250)]	0	[]
keras_layer (KerasLayer)	[(None, 768), (None, 250, 768)]	109482241	['input_word_ids[0][0]', 'input_mask[0][0]', 'segment_ids[0][0]']
tf.__operators__.getitem (SlicingOpLambda)	(None, 768)	0	['keras_layer[0][1]']
dense (Dense)	(None, 64)	49216	['tf.__operators__.getitem[0][0]']
dropout (Dropout)	(None, 64)	0	['dense[0][0]']
dense_1 (Dense)	(None, 32)	2080	['dropout[0][0]']
dropout_1 (Dropout)	(None, 32)	0	['dense_1[0][0]']
dense_2 (Dense)	(None, 4)	132	['dropout_1[0][0]']

Total params: 109,533,669
Trainable params: 109,533,668
Non-trainable params: 1

Figure 6.2 : TextBERT Model Summary

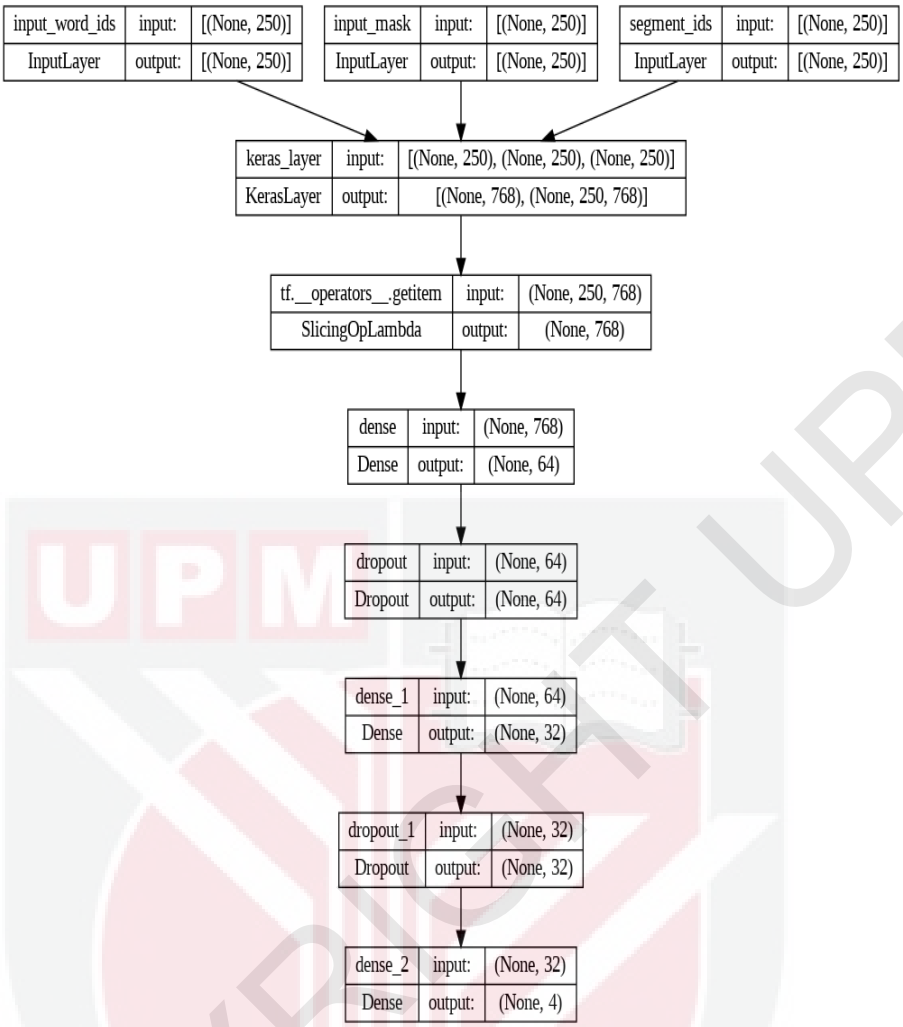

Figure 6.3 : Plot of TextBERT Model Graph

6.2.4 Proposed TextBERT Model Algorithm

The TextBERT model set of steps is outlined in Algorithm 3. Let D be the set of data, where: $D = \{(d_1, y_1), \ldots, (d_T, y_T)\}$. Here, (d_T, y_T), represents the feature set and label of a text data T. Tokenize each text input d in D into a sequence of sub words $w_1, w_2, \ldots, w_n$ using the BERT tokenizer. Then, encode each sequence of sub words $w_1, w_2, \ldots, w_n$ into a sequence of numerical inputs $x_1, x_2, \ldots, x_n$ using the BERT model's input encoding: $w_1, w_2, \ldots, w_n$ to $x_1, x_2, \ldots, x_n$. Next, add special tokens for classification; [CLS] and [SEP] to the beginning and end of the numerical input sequence: x_0 =

[CLS], x_1, x_2, ..., x_n, x_{n+1} = [SEP]. Pad or truncate the numerical input sequence to a fixed length, L: $x_{1:L}$ = (x_0, x_1, x_2, ..., x_n, x_{n+1}), where $n <= L - 2$. Following that, convert the numerical input sequence into input features: input_ids, input_mask and segment_ids).

Let B be the pre-trained BERT model. Add a classification layer on top of B to obtain a new model M_2. Train M_2 using the numerical input sequences $x_{1:L}$ and corresponding labels y in D. Tune the hyperparameters of M_2 to optimize performance. Once the optimal hyperparameters are found, retrain M_2 on the entire training set D. Lastly, evaluate the performance of the M_2 on the testing set and the final output will be the predicted class.

Algorithm 3: TextBERT Model	
Input:	$D = \{(d_1, y_1), \ldots, (d_T, y_T)\}$ $x_{1:L}$ = A fixed length of numerical input: (x_0, x_1, x_2, ..., x_n, x_{n+1}), where $n <= L - 2$.
Output:	$FinalOutput_{Test}$ = Label class predicted on the provided testing sample, $Test$ acc_{M2} = Percentage accuracy of M_2 fusion model M_2 represent model 2 which is the proposed TextBERT model.
1) Pre-processing: for each input text, d:	a) Tokenize the text into words, w and sub words, x b) Add special [CLS] and [SEP] tokens to the beginning, x_0 = [CLS] and end of each sentence, x_{n+1} = [SEP], respectively. c) Pad or truncate numerical input sequence to a fixed length, L: $x_{1:L}$ d) Convert the numerical input sequence into input features:(input_ids, input_mask and segment_ids) e) End for

2) Split D into training set (D_{train}), validation set (D_{val}) and testing set (D_{test})	
3) Load the pre-trained BERT model, B	
4) Add a dense classification layer on top B to create a model M_2	
5) Set hyperparameters; learning_rate, num_epochs, batch_size, Optimizer	
6) Train model: For epoch= 1 to num_epochs	a) Use D_{train} to get input and output tensors b) Pass the input tensors through TextBERT model, M_2 and compute output c) Compute loss and gradients d) Update model parameters e) Use D_{val} for validation and monitor performance e) End for
7) Compute the output of proposed TextBERT model using D_{test} for Ti, i = 1…nc (where nc = 4, the number of classes) to obtain the predicted label class, $FinalOutput_{Test}$	
8) Calculate the metrics such as accuracy, precision, recall, and F1 score	
9) Get the final percentage accuracy of proposed BERT model, acc_{M2}	
10) End	

Example Calculation for the Pre-Processing Component of Algorithm 3:

Suppose the input text is:

"I love you a great deal"

1. Tokenization

   Split the text into tokens (words and subwords) using a tokenizer:

   - Tokens: `['I', 'love', 'you', 'a', 'great', 'deal']`
   - Subwords (if applicable): `['I', 'love', 'you', 'a', 'great', 'deal']`

2. Add Special Tokens

   Add [CLS] at the beginning and [SEP] at the end:

   - Tokens: `['[CLS]', 'I', 'love', 'you', 'a', 'great', 'deal', '[SEP]']`

3. Pad or Truncate the Sequence

   Assume the fixed sequence length L=10L

   - If the token sequence is shorter than L, add [PAD] tokens.
   - Tokens after padding:

     `['[CLS]', 'I', 'love', 'you', 'a', 'great', 'deal', '[SEP]', '[PAD]', '[PAD]']`

4. Convert Tokens to Input Features

   Map the tokens to numerical IDs, create input masks, and segment IDs:

   - **input_ids:** `[101, 146, 2293, 2017, 1037, 2307, 3067, 102, 0, 0]`

     (Numerical IDs for each token)

   - **input_mask:** `[1, 1, 1, 1, 1, 1, 1, 1, 0, 0]`

     (Indicates actual tokens with `1` and padding with `0`)

   - **segment_ids:** `[0, 0, 0, 0, 0, 0, 0, 0, 0, 0]`

     (All 0 since this is a single-sentence input.)

For the input text "**I love you a great deal**", the pre-processed data is:

- **input_ids**: `[101, 146, 2293, 2017, 1037, 2307, 3067, 102, 0, 0]`
- **input_mask**: `[1, 1, 1, 1, 1, 1, 1, 1, 0, 0]`
- **segment_ids**: `[0, 0, 0, 0, 0, 0, 0, 0, 0, 0]`

This output is ready to be fed into the TextBERT model for further processing.

6.3 Evaluation and Result

This section discusses the evaluation procedure and its outcome of the proposed TextBERT model. This study used the text IEMOCAP dataset to evaluate the proposed TextBERT model. To maintain consistency with previous research, emotions of angry, happy, sad, and neutral were specifically selected from among all the categorical emotion labels that were annotated in the IEMOCAP dataset. 80%, 10% and 10% from the size of data were split for training validation and testing, respectively. The final size of training data is 3498 and testing data is 875. Table 6-1 shows the details of the dataset division after splitting.

Table 6.1 : Details of Dataset Division (Text)

IEMOCAP (Text) *(Involving 4 label class; angry, happy, sad, and neutral)*	
Number of Samples for Training Set	3498
Number of Samples for Validation and Testing Set	875

6.3.1 Hyperparameter Optimization for TextBERT Model

Fine-tuning a pre-trained TextBERT model involves replacing the last layer of the model with a new classification layer or dense layer and then training the model on a specific downstream task using a labelled dataset.

This study trains the entire architecture by using the dataset to train the complete pre-trained model and feed the results to a softmax layer. In this case, the error propagates backwards through the entire architecture and the model's pre-trained weights are updated according to the current dataset.

In order to fine-tune the TextBERT model, a few hyperparameters that would have a significant impact on the model's final performance were considered and taken into account. Table 6-2 summarizes the tuned hyperparameter values for the TexBERT model.

Table 6.2 : Hyperparameters Tuning Value (TextBERT)

Used Hyperparameter	Values
Number of Epochs	100
Batch Size	64
Learning Rate	2e-5
Dropout Rate	0.2
Optimizer	Adam

6.3.2 Experimantal Set Up for TextBERT Model

To evaluate the effectiveness of the proposed TextBERT model, an experimental design was set up to compare the performance between the proposed TextBERT method with other state-of-the-art methods in the text emotion recognition. The state-of-the-art methods involved have also been trained using the same dataset from IEMOCAP text data.

6.3.3 Result

The evaluation result and analysis of the TextBERT model are presented in this section. Figure 6-4 shows the accuracy per epoch and loss per epoch of the TextBERT model.

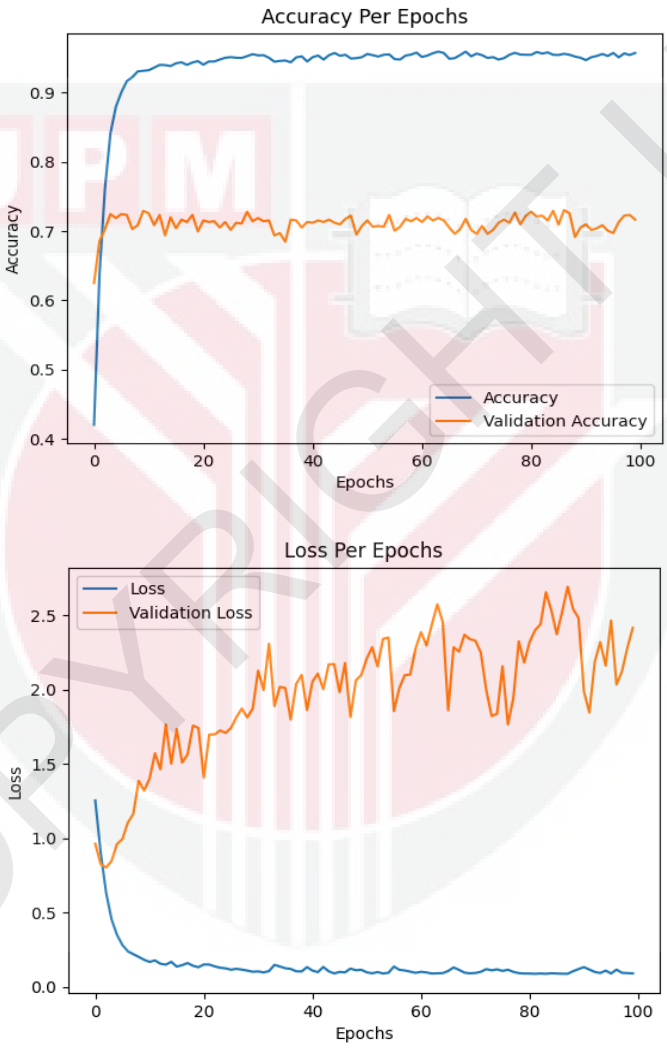

Figure 6.4 : Accuracy Per Epoch and Loss Per Epoch of TextBERT Model

As shown in Figure 6-5, the TextBERT model achieves 70% accuracy with loss = 2.5. Based on the confusion matrix shown in Figure 5-5, TextBERT model achieves high True Positive Rate in most of the emotions such as angry, happy, sad with their

corresponding TPR = 81%, 74%, 73%, whereas the model performs slightly less effectively on neutral emotion with TPR = 54%.

Figure 6.5 : Confusion Matrix of TextBERT Model

Table 6-3 presents the result of precision, recall and F1 score of each emotion label; angry, happiness, sadness and neutral for the TextBERT model. The happy emotion yields the highest precision with 76%. For recall, angry emotion yields the highest score with 81%. Meanwhile for F1-score, happy emotion once again achieves the best score with 75%. This study observes that the high precision for happy suggests the model effectively avoids false positives, likely due to the well-defined and distinct nature of happiness in the dataset. The high recall for angry reflects the model's ability to accurately identify almost all instances of anger, aided by clear textual patterns and BERT's bidirectional processing. The strong F1-score for "happy" indicates that the model consistently balances both precision and recall for this emotion. These results

are influenced by the dataset's clear examples of happy and angry emotions, as well as BERT's contextual embeddings, which capture subtle emotional cues.

Table 6.3 : Precision, Recall and F1-Score of TextBERT Model

Emotion	Precision (%)	Recall (%)	F1-Score (%)
Angry	69	**81**	74
Happy	**76**	74	**75**
Sad	70	73	71
Neutral	65	54	59

Table 6-4 and Figure 6-6 provide a comparative analysis of percentage accuracy, including a graphical representation for state-of-the-art methods and the proposed TextBERT model in text-based emotion recognition. To ensure consistency with previous studies, both unweighted accuracy (UA) and weighted accuracy (WA) were recorded. As shown in Table 6-4, the TextBERT model achieves the highest UA and WA, both at 70%. This represents a notable improvement of approximately 5% over the best-performing state-of-the-art method.

Table 6.4 : Comparison TextBERT Model with State-of-The-Art

Method	UA (%)	WA (%)
LSTM + Attn (Xu et al., 2019)	57.8	63.3
Dual LSTM (Tripathi et al., 2019)	58.4	64.7
2Transformers Encoder (J. Zhang et al., 2022)	61.4	66.0
TextBERT	70.0	70.0

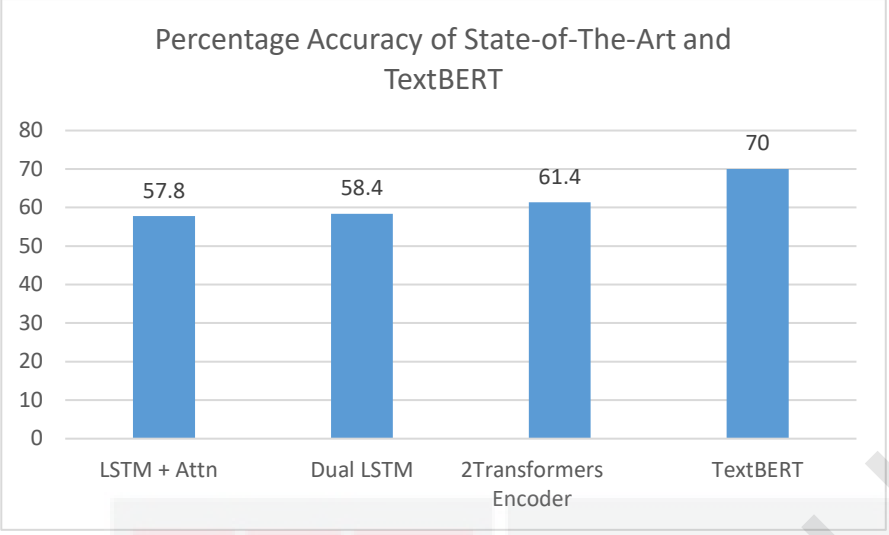

Figure 6.6 : Percentage Accuracy of State-of-The-Art and TextBERT Model

6.4 Analysis and Discussion

This study implemented the BERT language model for the development of text emotion recognition model. TextBERT has demonstrated outstanding results in the area of emotion recognition, exceeding other cutting-edge text emotion recognition models. The TextBERT model can accurately capture the context of a text because of the BERT language model's bi-directional nature, which considers information from both the left and right sides of a word. TextBERT acquires a comprehensive representation of language that enables it to identify emotions according to the context meaning.

TextBERT model's uniqueness lies in its customization for lightweight model strategies and domain adaptation, specifically for emotion recognition. While the BERT architecture itself has been widely used in various natural language processing tasks, the proposed TextBERT model's distinctiveness comes from how it has been tailored to address specific requirements within the emotion recognition domain.

Firstly, by incorporating a classification layer with two dense layers and employing an embedding dimension of 768, the TextBERT model has been optimized for computational efficiency. The classification layer, together with two additional dense layers on top of the pre-trained BERT embeddings, results in a more compact model that preserves BERT's essential capabilities while reducing its computational burden. It enables the TextBERT to deliver faster predictions and consume fewer resources, making it more suitable for deployment in various scenarios. Furthermore, the selection of the embedding dimension reflects the TextBERT model's ability to extract more detailed contextual information, which is useful for tasks that require a more complex comprehension of the text.

Secondly, the integration of dropout regularization in the TextBERT model is a strategic move. Dropout can help mitigate overfitting, enhancing the generalization capabilities of the model. By applying dropout within two dense layers, it is effectively preventing the network from relying too heavily on any single feature thus, making it more robust and reliable. Besides that, the strategy used allows TextBERT to adapt more effectively to the nuances of different emotional contexts. This domain adaptation capability enables TextBERT to excel in recognizing emotions in a variety of domains or user-specific scenarios, making it highly versatile and valuable for applications.

6.5 Summary of Chapter 6

This chapter introduces TextBERT, a text emotion recognition model developed using the BERT (Bidirectional Encoder Representations from Transformers) framework. The primary objective is to address the challenges in recognizing emotions from

human speech text data, which often include indirect expressions of emotions and contextual ambiguities. TextBERT aims to capture the context-dependent relationships between words. By optimizing the model's architecture, the study presents a lightweight version of BERT that balances computational efficiency with high performance in emotion recognition tasks.

The architecture of TextBERT is designed to optimize performance for emotion recognition tasks. A dense classification layer is added on top of BERT embeddings, which are fine-tuned using the IEMOCAP dataset. The evaluation of TextBERT on the IEMOCAP dataset demonstrates its effectiveness in recognizing emotions like angry, happy, sad, and neutral. The model achieved a classification accuracy of 70%, outperforming several state-of-the-art methods in text emotion recognition.

In conclusion, TextBERT has exhibited improved potential in the area of text emotion recognition, reaching high accuracy and exceeding other cutting-edge text emotion recognition models. The pre-training approach of the BERT language model has shown to be a successful way to extract context information from text data, and its ability to be fine-tuned makes it simple to adapt to an emotion recognition task. Furthermore, the TextBERT model distinguishes itself as a versatile and efficient solution that aligns to the lightweight model strategy and is well-equiped for domain adaptation tasks, where it can efficiently adapt to different datasets and tasks while maintaining robust performance.

CHAPTER 7

Fer2DCNN FOR VIDEO IMAGE EMOTION RECOGNITION MODEL

7.1 Introduction

In this chapter, the focus was given to video modality of data. The main objective of this chapter is to extract deep features from video data and develop an emotion recognition model to recognize human's emotion. A video is a series of images referred to as frames captured over a period of time. Frame per second is an essential feature in video as it indicates how many images are taken in a second. Figure 7-1 illustrates how each image in the video changes over time while maintaining the same size and quality.

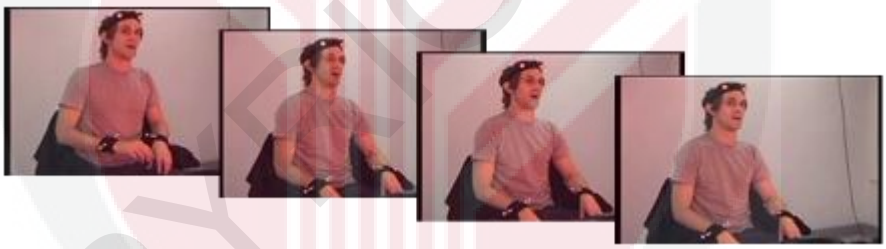

Figure 7.1 : Video Frames

Overfitting issues often occur in video classification. This happens when in most cases related to video classification, only a small part of the video segment is highly related to the video content while the majority of the rest are not related. Besides that, a video contains a lot of emotion. Finding the emotion that truly dominates and has the greatest impact on the interpretation of a video can be challenging. Therefore, to address this issue, this study focused a lot of work and attention on the video pre-processing part and also carefully extracted the important and relevant features to be included in the training model. This study performs data augmentation to obtain a clear view image

sample from the video data. Besides that, the Cascade Haar classifier from the OpenCV library was implemented to detect the facial expression of the actors from the video frames.

Spatial and temporal are very important elements in video classification. Many researchers have implemented 3DCNN in their work especially in emotion recognition. Despite their successes, 3DCNN has a number of drawbacks that limit their performance and usability. Training and testing with 3DCNN can be very challenging due to the high computational cost and the fact that 3DCNN training requires a lot of parameters and computations. In terms of emotion recognition tasks, 3DCNN are prone to overfitting, especially when trained on small datasets. Often, there is limited availability of annotated video data for emotion recognition tasks, making it difficult to train and validate 3DCNN models. The models may become less resistant to changes in the distribution of emotions as a result of overfitting, which can also result in poor generalization performance.

This study introduced Fer2DCNN, a customized 2DCNN algorithm with an advanced pre-processing approach with the emphasis on facial expression detection. The data used is video data from the IEMOCAP dataset. After carefully analyzing the data, the actors in the video were found to have mostly kept their appearance over time and that the only noticeable changes are in facial expressions. In order to capture the changing facial expressions in the video data, individual video frames from the video were extracted and processed as individual images. Further description regarding the proposed system will be presented in the following sections of this chapter.

7.2 Video Data Pre-Processing

This section presents in detail on IEMOCAP video data pre-processing which is an important step before training models and has a significant impact on the accuracy of the final results. Each video data in the IEMOCAP dataset was sliced into sentences in accordance with how the audio data was processed. This step was taken to ensure the access part of the video corresponds with the provided audio data in the IEMOCAP dataset.

7.2.1 Frame Extraction

As mentioned earlier, video data consists of a series of still images. Extraction of frames from a video in computer vision can be beneficial for a number of tasks, including object detection, tracking, and face recognition. In this study, OpenCV was used to extract frames from video data at a given time. The extraction of frames from a video using OpenCV and Python is a straightforward task that can be completed in a few simple steps.

First, the required libraries and packages such as OpenCV and Numpy libraries must be imported. Next, the cv2.VideoCapture function was used to load the video file. This function accepts the path to the video file as a parameter and returns a VideoCapture object. Once the video files have been loaded, the frames from video using cap.read() function will be then extracted. The cap.read() function returns a tuple of two values whereby the first value is a Boolean that indicates whether or not the frame was successfully read while the second value is the frame itself.

From the entire duration of the sliced video eight frames were extracted and saved as images. The extracted frames are displayed using the cv2.imshow() function. The cv2.waitKey() function was used to wait for a key event. The loop will end if the "q" key is pressed. After all the frames have been extracted, the video file was released and all the windows were destroyed. There are two actors that were initially visible in every video and because of that, each frame produced was then cropped accordingly to capture only the actor whose emotion is being recorded and minimize the unwanted background (refer Figure 7-2). Finally, the emotion class of each cropped frame was identified and saved in the respective folder.

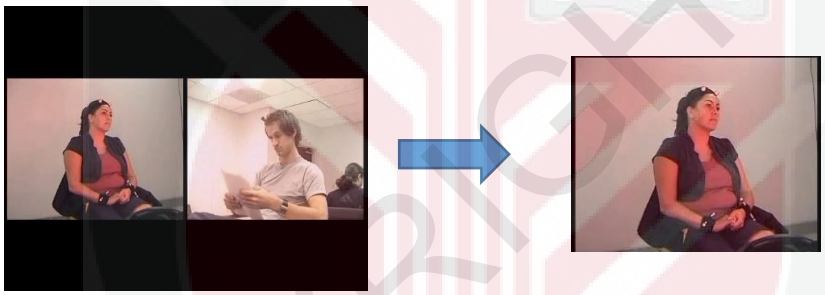

Figure 7.2 : Example of Cropped Frame

7.2.2 Face Detection with Haar Cascade Classifier

The Haar Cascade Classifier is an established technique used for object detection in images and videos. It was first presented by Viola and Jones in 2001 and has since grown to be one of the most popular object detection techniques in computer vision (Viola & Jones, 2001).

A. Theoretical of Haar Cascade Classifier

The Haar Cascade Classifier employs a number of Haar features, which are simple rectangular and angled shapes used to detect edges, lines, and textures in an image (refer Figure 7-3). They provide the same function as a convolutional kernel. The system calculates the total pixel intensities in the rectangle's white area and subtracts it from the total pixel intensities in the black area. The value obtained serves as a feature to denote the existence of an object.

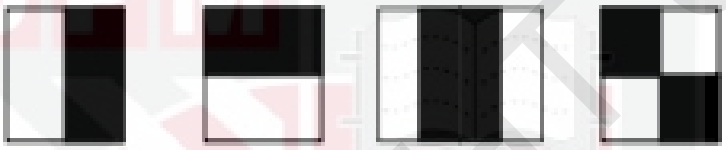

Figure 7.3 : Illustration of Haar-Like Features Used by Viola and Jones in 2001

The Integral Image is a concept used in the Haar Cascade Classifier to simplify the calculation of Haar-like features and speed up the process of object detection. By using the Integral Image, the classifier is able to calculate the total of intensity values inside a certain area without having to go through the full image. The intensity levels of the original image are added together and it is determined by adding up all the pixels from the image's top-left corner to the current location. As a result, a new image is created with the same height and width as the original image but with the pixel values representing the total of all the pixels above and to the left of that point.

For Haar Cascade Classifiers to be more accurate at detecting objects of interest, AdaBoost is frequently implemented. The weak classifier in this case is the Haar Cascade Classifier, and the AdaBoost algorithm is used to integrate multiple weak classifiers to form a strong classifier. The AdaBoost algorithm adjusts the weights of

the weak classifiers to ensure that the overall classifier is more accurate in detecting objects of interest.

Another important element in the Haar Cascade Classifiers is Attentional Cascade. The attentional cascade concept in Haar cascade classifiers refers to a series of stages where each stage acts as a filter to eliminate non-face regions, focusing on only the regions of interest. In this case, a window will slide across image areas that will first be labelled as positive or negative. The window will slide to the subsequent region if the region has a negative label. The classifier will move on to the next stage if the result is positive. At the final stage, the object is considered found if the label is still positive.

B. Implementation of Haar Cascade Classifier

The Cascade Haar classifier is a straightforward yet effective method for face detection. This method can be easily implemented in the Python programming language using OpenCV library. In order to use the Cascade Haar Classifier for face detection in OpenCV, import the pre-trained classifier from the XML file into the program. This file contains the learned parameters of the classifier which can be loaded using the cv2.CascadeClassifier function.

Next, the cv2.imread() function was used to load the input image from video frames data that will be required for face detection. The Haar classifier requires grayscale images to function effectively. It is an optional but for the Haar classifier to function effectively, the image should be in grayscale format. The converting process of the image to grayscale can be done using the cv2.cvtColor() function. However, in this

case, the colour of images was chosen to remain as they are without converting them into grayscale.

Finally, the faces in the image were detected using the detectMultiScale() function. This function accepts a grayscale image as input and outputs a list of faces found in the image. The cv2.rectangle() function was implemented to draw rectangles around the faces and the cv2.imshow() function was then used to display the output image with the rectangles. The final resolution for each image has been set to 64 (L) x 64 (W).

Figure 7-4 shows the face detection on the image data after applying the Haar Cascade Classifier.

Figure 7.4 : Face Detection after Applying Haar Cascade Classifier

7.3 Proposed Model of Fer2DCNN

This section describes in detail about the proposed Fer2DCNN of image classification model for emotion recognition. The proposed design of the Fer2DCNN model is illustrated in Figure 7-5.

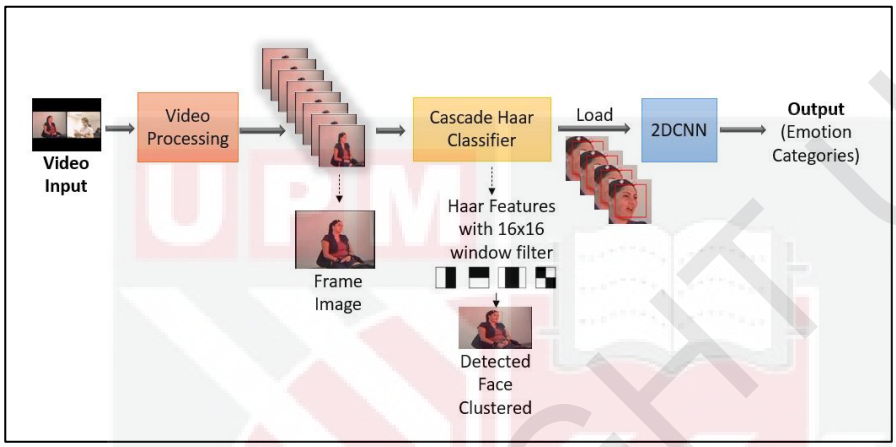

Figure 7.5 : Illustration of Proposed Fer2DCNN

The Fer2DCNN model is designed for emotion recognition by analyzing facial expressions from video data, particularly using the IEMOCAP dataset. The process begins with frame extraction, where all frames from a video are treated as individual images. Each frame undergoes face detection using the cascade Haar classifier, a robust method that isolates the face region while ignoring irrelevant background elements. This ensures the model focuses solely on the facial features that are crucial for emotion recognition.

Once the face is detected, the individual frames are fed into the 2D Convolutional Neural Network (2DCNN). Unlike methods that use only a single frame for classification, Fer2DCNN processes every frame in the video. This approach allows the model to analyze facial expressions over time, capturing the dynamic progression

of emotions within the video. By examining all frames, the model can detect even subtle changes in expressions, providing a more comprehensive understanding of the emotional state.

For labelling, a straightforward method is to assign the same label to all frames in a video clip if the emotion remains consistent throughout the clip. For instance, if a video corresponds to a specific emotional state such as happiness, sadness, or anger, every frame extracted from the video will share that same label. This method is particularly effective in controlled environments where actors maintain a consistent facial expression throughout the clip

The 2DCNN then extracts features from each detected face using convolutional layers, which identify patterns related to facial expressions, and pooling layers, which emphasize the most critical features. Finally, the model classifies the emotion for each frame based on predefined categories. By processing the entire sequence of frames, the model enhances its ability to recognize emotions accurately, accounting for temporal variations in facial expressions.

7.3.1 Fer2DCNN Detailed Description

Fer2DCNN receives input images that focus on the part of the facial expression which has been extracted from the video data. The 2DCNN extracts features from the input image using a series of filters known as convolutional layers. These filters are applied to specific areas of the image, allowing the network to recognize patterns and spatial relationships between pixels. Besides that, these filters also detect specific features of

the image like corners or edges and combine them to produce a feature map. The convolution process can be expressed in the following formula:

$$(x * h)_{i,j} = \sum_{m}^{k_1-1} \sum_{n}^{k_1-2} x_{i+m,j+n}.h_{m,n} + b$$

$$0 \leq i \leq n_1 - 1 \; and \; 0 \leq j \leq n_1 - 1$$

(Equation 31)

where, *x* denotes the input vector for this convolutional layer, *h* denotes the filter and *b* is the constant bias.

Moving forward to the pooling layer, when given an image $x(k1, k2)$ and a stride of *k*, the max function, *f*max produces an output image in which each *k x k* block is reduced to a single value. *f*max returns the maximum value from $x(k1, k2)$ for each *k x k* block.

$$Px_{ij} = \max \begin{bmatrix} x_{ki,kj} & x_{ki,kj+1} & \cdots & x_{ki,kj+k-1} \\ x_{ki+1,kj} & x_{ki+1,kj+1} & \cdots & x_{ki+1,kj+k-1} \\ \vdots & \vdots & \cdots & \vdots \\ x_{ki+k-1,kj} & x_{ki+k-1,kj+1} & \cdots & x_{ki+k-1,kj+k-1} \end{bmatrix}$$

$$0 \leq i \leq \left[\frac{n_1 - 1}{k}\right] \; and \; 0 \leq j \leq \left[\frac{n_2 - 1}{k}\right]$$

In the fully connected layer, the two main operations that take place are forward propagation and back propagation. In forward propagation, each neuron in the layer receives inputs from all the neurons in the previous layer, multiplies them by their respective weights, sums them up, and applies an activation function to produce an output. The output of a neuron in the fully connected layer can be presented mathematically as:

$$z = w1 * x1 + w2 * x2 + \ldots + wn * xn + b$$

(Equation 32)

where, w1, w2, ..., wn are the weights of the connections between the neuron and the input neurons, x1, x2, ..., xn, are the inputs, b is the bias, and z is the output of the neuron.

The output of the entire layer can be represented as a vector of outputs of all neurons in the layer:

$$Z = [z1, z2, \ldots, zm]$$

where m is the number of neurons in the layer.

The final output of the layer is then produced by passing the output vector Z through an activation function, such as a sigmoid or ReLU function. This process can be described by the equation below:

$$y = f(Z)$$

(Equation 33)

Meanwhile, the back-propagation process involves figuring out the gradients of the error with respect to the weights and biases of the layer. The weights and biases in the layer are then updated using these gradients in order to reduce error in the subsequent iteration. The chain rule of calculus can be used to determine the gradients:

$$\frac{\delta E}{\delta w} = \frac{\delta E}{\delta z} * \frac{\delta z}{\delta w}$$

$$\frac{\delta E}{\delta b} = \frac{\delta E}{\delta z} * \frac{\delta z}{\delta b}$$

(Equation 34)

where, E is error, z is output of neuron, w is weight and b are bias.

The following formulas can be used to calculate the gradients of the error with regard to the layer's output:

$$\frac{\delta E}{\delta z} = \frac{\delta E}{\delta y} * \frac{\delta y}{\delta z}$$

(Equation 35)

The gradient descent algorithm is then used to update the weights and biases:

$$w = w - \propto * \frac{\delta E}{\delta w}$$

$$b = b - \propto * \frac{\delta E}{\delta b}$$

(Equation 36)

where, $\propto$ is learning rate / the step size used to update the weights and biases.

Both forward and back propagation processes will repeat alternately until the error is equal to the targeted minimum error.

Apart from that, the categorical cross-entropy with integer targets also known as sparse categorical cross-entropy loss function was employed to train the proposed model. The sparse categorical cross-entropy has the same loss function with categorical cross-entropy. The only difference between them is in the way labels are defined. In sparse categorical cross-entropy, labels are integer encoded, for instance

[1], [2], and [3] for a 3-class problem. The usage of sparse categorical cross-entropy has the benefit of saving both memory and calculation time by only requiring a single integer to represent a class rather than an entire vector.

7.3.2 Fer2DCNN Model Architecture

The architecture of the proposed Fer2DCNN network is presented as follows:

- **CONV 1:** The input shape of the first convolutional layer is 64 x 64 x 3. This layer implements 32 kernels with spatial size of 3 x 3. The activation function used is the Leaky Rectified Linear Unit (ReLU). Followed by a 2 x 2 stride step max-pooling function with Batch Normalization.
- **CONV 2:** The second layer implements 64 kernels with spatial size of 3 x 3. The activation function used is the Leaky Rectified Linear Unit (ReLU). Followed by a 2 x 2 stride step max-pooling function.
- **DENSE 1:** The first dense layer sets 128 units with the implementation of Leaky Rectified Linear Unit (ReLU) as the activation function.
- **DENSE 2:** The last dense layer sets 4 units with the implementation of Softmax as the activation function.

Figure 7-6, 7-7 and 7-8 present the proposed Fer2DCNN model architectures, the model summary and plot of proposed 2DCNN model graph, respecti

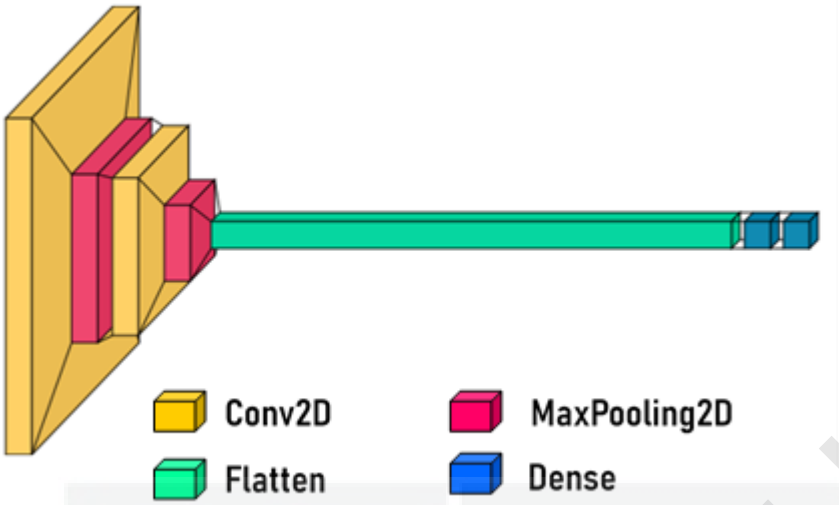

Figure 7.6 : Proposed Fer2DCNN Model Architectures

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 62, 62, 32)        896

max_pooling2d(MaxPooling2D)  (None, 31, 31, 32)        0

conv2d_1 (Conv2D)            (None, 29, 29, 64)        18496

max_pooling2d_1              (None, 14, 14, 64)        0
(MaxPooling2D)

flatten (Flatten)            (None, 12544)             0

dense (Dense)                (None, 128)               1605760

dense_1 (Dense)              (None, 4)                 516

=================================================================
Total params: 1,625,668
Trainable params: 1,625,668
Non-trainable params: 0
_________________________________________________________________
```

Figure 7.7 : Proposed Fer2DCNN Model Summar

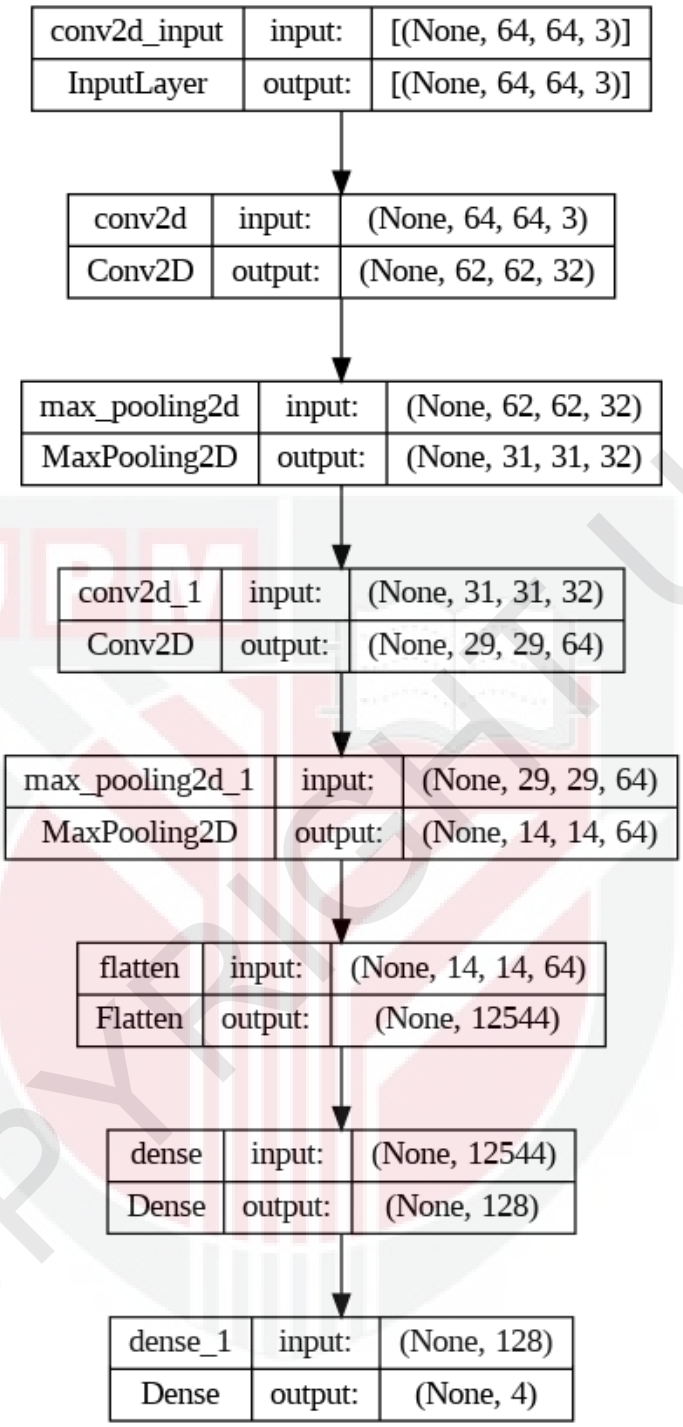

Figure 7.8 : Plot of Proposed Fer2DCNN Model Graph

7.3.3 Fer2DCNN Model Algorithm

The proposed Fer2DCNN model set of steps is outlined in Algorithm 4. Let D be the set of data, where: $D = \{(d_1, y_1), \ldots, (d_I, y_I)\}$. Here, (d_I, y_I), represents the feature set

and label of the image *I* extracted from video data. Each image frame was looped to detect faces by using the Haar Cascade Classifier. Let {x, y, w, h} represent the edges of the detected face region. Each detected face was then looped to draw a rectangle around it. The face region was resized to a fixed size (64 x 64) and saved as separate images, *x*. Besides that, the augmentation to the face region was applied before being fed into the 2DCNN.

The data were split into training, validation and testing sets with an 80:10:10 ratio. The 2DCNN model, M_3 was then constructed by specifying its architecture, incorporating convolutional and max pooling layers, flattening the output of the convolutional layers, adding dense layers, and applying an appropriate softmax activation function for classification purposes. The model was compiled by selecting an appropriate loss and optimizer and trained on a labelled training dataset using backpropagation and gradient descent. All the hyperparameters were tuned to optimize the performance. Finally, the model's performance was assessed by using the testing dataset, and the predicted class was considered to be the final output.

Algorithm 4: Fer2DCNN Model	
Input:	$D = \{(d_1, y_1), \ldots, (d_I, y_I)\}$ $(x_1, \ldots x_I)$= Images on the face regions
Output:	$FinalOutput_{Test}$ = Label class predicted on the provided testing sample, *Test* acc_{M3} = Percentage accuracy of M_1 fusion model M_3 represent model 3 which is the proposed Fer2DCNN.
1) Loop through image frames: For d = 1 to I	a) Detect faces using the Haar Cascade Classifier b) Loop through detected face region: For (x, y, w, h) in *d*:

	i) Draw a rectangle around the face ii) Resize the face region to a fixed size (64 x 64) iii) Save the faces as separate images, *x* iv) End For c) End For
2) Split *D* into training set (D_{train}), validation set (D_{val}) and testing set (D_{test})	
3) Define 2DCNN Model architecture	
4) Initialize weights and biases randomly	
5) Set hyperparameters; learning_rate, num_epochs, batch_size, optimizer	
6) Train model: For epoch= 1 to num_epochs	a) Loop over batches of data, D_{train} b) Extract batch of input data and corresponding labels from D_{train} c) Perform a forward pass using Equation 31,32 and 33 d) End For
7)Perform backward pass: For each layer *l* in reverse order (from output to input)	a) Compute the gradient of the output w.r.t. the input for layer *l* using equation 34. b) Compute gradients of the loss w.r.t. weights and biases for layer *l* using equation 35. c) Update the weights and biases for layer *l* using equation 36 and the optimizer. d) Update the gradients for the previous layer. e) End For
9) Compute the output of proposed TextBERT model using D_{test} for *Ti*, $i = 1 \ldots nc$ (where $nc = 4$, the number of classes) to obtain the predicted label class, $FinalOutput_{Test}$	
10) Calculate the metrics such as accuracy, precision, recall, and F1 score	
11) Get the final percentage accuracy of proposed Fer2DCNN model, acc_{M3}	
12) End	

This example demonstrates the flow of Algorithm 4, showing how the Fer2DCNN model processes data through training, validation, and testing stages:

1. Preprocessing (Frame Extraction and Face Detection)
    - Suppose we have a video with 10 frames ($D=\{d_1,d_2,\ldots,d_{10}\}$
    - Each frame undergoes Haar Cascade face detection to extract the face region.
    - The detected face is cropped and resized to 64×64
    - After preprocessing, we now have 10 face images: $X= \{x_1, x_2\ldots, x_{10}\}$

2. Splitting the Data
    - The processed data is split into:
        - Training set (D_{train}): 8 images ($x_1,\ldots,x_8$).
        - Validation set (D_{val}): 1 image (x_9).
        - Testing set (D_{test}): 1 image (x_{10}).

3. Training the Model
    - For the first training batch (x_1, x_2):
        - Forward Pass:
            - The model predicts probabilities. For example:

              For x_1: [0.2,0.1,0.6,0.1] → "happy".

              For x_2: [0.3,0.5,0.1,0.1] → "neutral".
        - Loss Calculation:
            - For x_1 (true label = "happy"):
              $-\log(0.6) = 0.51$

- For x2 (true label = "neutral"):
  $-\log(0.5) = 0.69$

- Weight Updates: The model adjusts weights using backpropagation to minimize the loss.

- The training continues for all batches and epochs, refining the model's performance.

4. Validation

- After training, the model's performance is validated on the validation set (D_{val}).
  - For x_9, predicted probabilities might be [0.2,0.3,0.4,0.1] → "sad".
  - The validation loss and accuracy are calculated to tune hyperparameters and prevent overfitting.

5. Testing Model

- The model is then tested on $D_{test}=\{x_{10}\}$
  - For x_{10}, predicted probabilities might be [0.3,0.1,0.1,0.5] → "neutral".

6. Evaluate Metric

7.4 Evaluations and Results

This section discusses in detail regarding the evaluation procedure and result of the proposed Fer2DCNN. The data used is video data from the IEMOCAP dataset. To maintain consistency with previous research, the emotions of angry, happy, sad, and neutral were specifically selected from among all the categorical emotion labels that were annotated in the IEMOCAP dataset. To ensure that all labels are balanced, each label was ensured to have the same amount of data. 80%,10% and 10% from the size

of data were split for training, validation and testing, respectively. From the 4339 number of samples data, 3472 have been assigned for training while 867 for testing. Table 7-1 shows the details of the dataset division after splitting.

Table 7.1 : Details of Dataset Division (Video)

IEMOCAP (Video) *(Involving 4 label class; angry, happy, sad, and neutral)*	
Number of Samples for Training Set	3472
Number of Samples for Validation and Testing Set	867

Despite the fact that the video data initially required a focus on spatiotemporal features for its classification, after carefully examining the video data, the variations in the motions and facial expressions are not very noticeable from one frame to the other frame. In order to optimize the data dimensionality and lower the chance of overfitting, focus was given on spatial rather than on temporal features.

7.4.1 Hyperparameter Optimization for Fer2DCNN

This study aims to identify the best set of hyperparameters that will maximize the model accuracy on the testing set. A few hyperparameters were considered and tuned for the proposed Fer2DCNN. Table 7-2 presents the tuned hyperparameter values for the proposed Fer2DCNN model.

Table 7.2 : Hyperparameters Tuning Value Fer2DCNN

Used Hyperparameter	Values
Number of Epochs	50
Batch Size	32
Learning Rate	1e-3
Kernel	3
Optimizer	Adam

7.4.2 Experimental Set Up for Fer2DCNN Model

In this section, a thorough evaluation set up consisting of three experimental designs was developed to assess the efficacy of the proposed Fer2DCNN model.

To ensure that the Fer2DCNN model is fairly compared to the state-of-the-art 3DCNN model, the first experimental design involves evaluating the performance of Fer2DCNN against a standard 3DCNN model. In this experiment, two different feature input sets were used, generated according to the pre-processing steps outlined earlier. The Fer2DCNN model emphasizes facial expressions by applying the Haar Cascade Classifier to individual frames sampled from the original sliced video data. In contrast, the 3DCNN model does not prioritize facial expressions in this way. Instead, the 3DCNN was trained and tested with sequences of eight frames from each sliced video, where the facial expression emphasis is not applied. The results of both the Fer2DCNN and 3DCNN models were then compared based on the percentage accuracy of emotion classification. This comparison clarifies the differences in approach between the two models, highlighting the unique design choices that may contribute to the observed performance differences.

The second experimental design involves comparing the performance of the proposed Fer2DCNN method with other state-of-the-art methods in the visual classification model for emotion recognition. Here, visual classification refers to the use of visual stimuli such as video, image, or motion capture data. The state-of-the-art methods compared include both 2D CNN and 3D CNN-based models, such as 3DResNet and 3DRSN, which have also been trained using the same IEMOCAP video dataset.

7.4.3 Result

The evaluation result and findings of the proposed system based on the aforementioned sets of experimental design are presented in this section.

A. Experiment 1: Comparison between Fer2DCNN and 3DCNN

The first experiment was conducted to compare the performance of the proposed Fer2DCNN with 3DCNN model. Table 7-3 shows the experiment results of the percentage accuracy of the proposed Fer2DCNN model and 3DCNN model

Table 7.3 : Comparison of Different Form of Input Feature Set

Model	Accuracy (%)
Fer2DCNN	84
3DCNN	63

Based on the result, the proposed Fer2DCNN model achieves higher accuracy with 84% accuracy, which outperforms the result of 3DCNN model with 63% accuracy. It shows that the proposed model is 21% significantly higher than the accuracy achieved by the 3DCNN model. Figure 7-9 illustrates the accuracy per epoch and loss per epoch of the proposed Fer2DCNN.

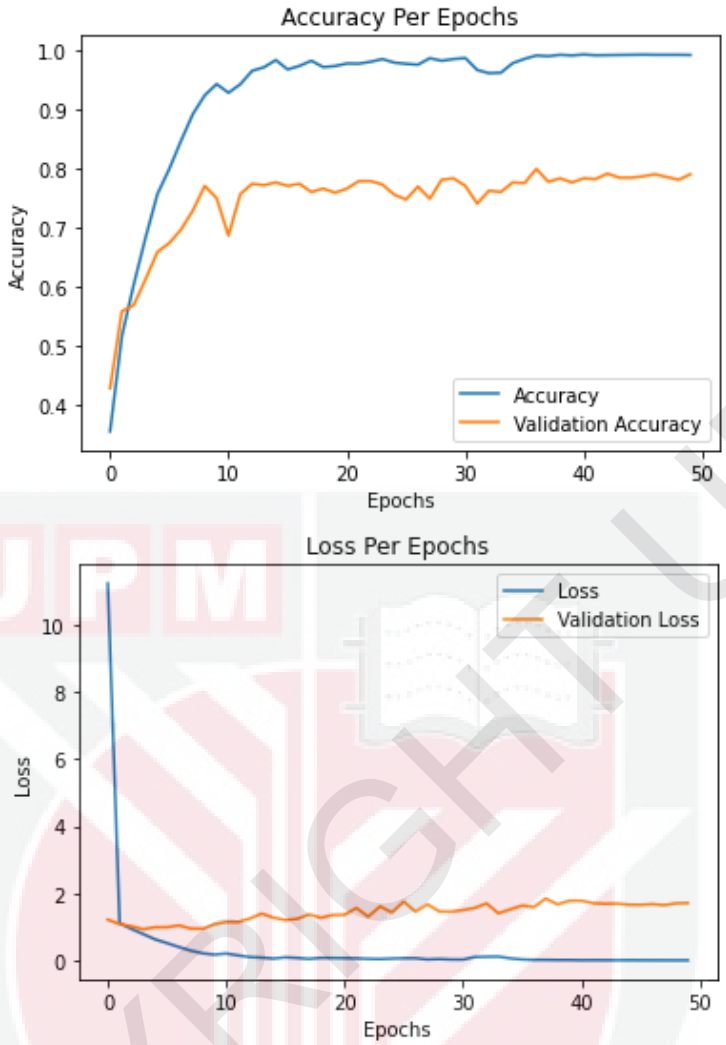

Figure 7.9 : Accuracy Per Epoch and Loss Per Epoch of Fer2DCNN Model

Based on the confusion matrix shown in Figure 7-10, Fer2DCNN model achieves high True Positive Rate in most of the emotions such as happy, sad and neutral with their corresponding TPR = 92%, 78%, 100%, whereas the model performs slightly less effectively on angry emotion with TPR = 67%.

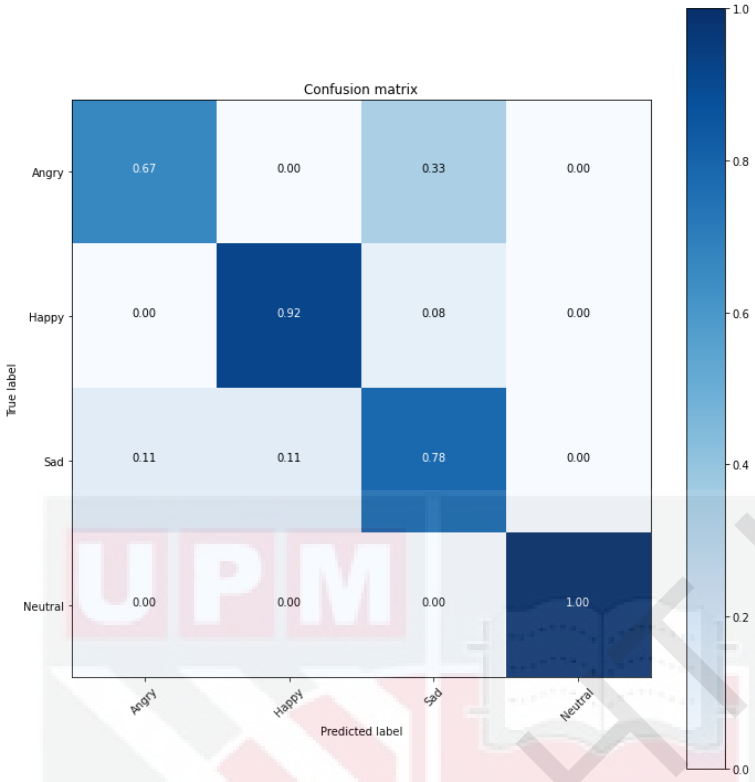

Figure 7.10 : Confusion Matrix of Fer2DCNN Model

Table 7-4 shows the result of precision, recall and F1 score of each emotion label; angry, happiness, sadness and neutral. Among all, neutral emotion scores highest precision recall and F1-Scores with 100%. On the other hand, this study observes that the lower performance of the angry emotion in terms of Precision**,** Recall**,** and F1-Score in Table 7-4 can be attributed to several factors. First, emotions like angry often overlap with other emotions such as sad or neutral, particularly in facial expressions and tonal variations. This overlap can lead to confusion during classification, resulting in misclassifications and reduced performance metrics. Furthermore, the angry emotion exhibits greater variability in expressions, which makes it challenging for the model to generalize and detect consistent patterns.

Table 7.4 : Precision, Recall and F1-Score of Fer2DCNN

Emotion	Precision (%)	Recall (%)	F1-Score (%)
Angry	80	67	73
Happy	92	**92**	**92**
Sad	**70**	78	74
Neutral	100	100	100

B. Experiment 2: Comparison with State-of-The-Art

Table 7-5 and Figure 7-11 present a comparison of percentage accuracy and graphical representation for state-of-the-art methods and the proposed Fer2DCNN in the video or image classification model for emotion recognition. The un-weighted accuracy (UA) and weighted accuracy (WA) were considered in the reported percentage accuracy results to maintain consistency with prior studies on the state-of-the-art methods. As shown in Table 7-5, the proposed Fer2DCNN achieves the highest percentage of UA and WA, both at 84%. This marks a significant improvement of approximately 30% over the best-performing method among the other state-of-the-art models.

It is important to note that, as shown in Table 7-3, the standard 3DCNN model used in this study achieves a result that is 12% higher than the 3DResNet and 3DRSN models, which have been claimed as state-of-the-art. These models, although based on 3DCNN architectures, might show performance differences due to factors such as dataset characteristics, training configurations or specific architectural details. The observed performance discrepancy does not imply that the compared models are not "state-of-the-art," but rather highlights the potential advantages of the specific approach, configuration and training strategies used in the proposed Fer2DCNN model.

Table 7.5 : Comparison Fer2DCNN with State-of-The-Art

Method	UA (%)	WA (%)
CNN (Tripathi et al., n.d.)	45.3	51.1
3Resnet (J. Zhang et al., 2022)	51.2	55.4
3DRSN (J. Zhang et al., 2022)	51.4	56.2
Fer2DCNN	84.0	84.0

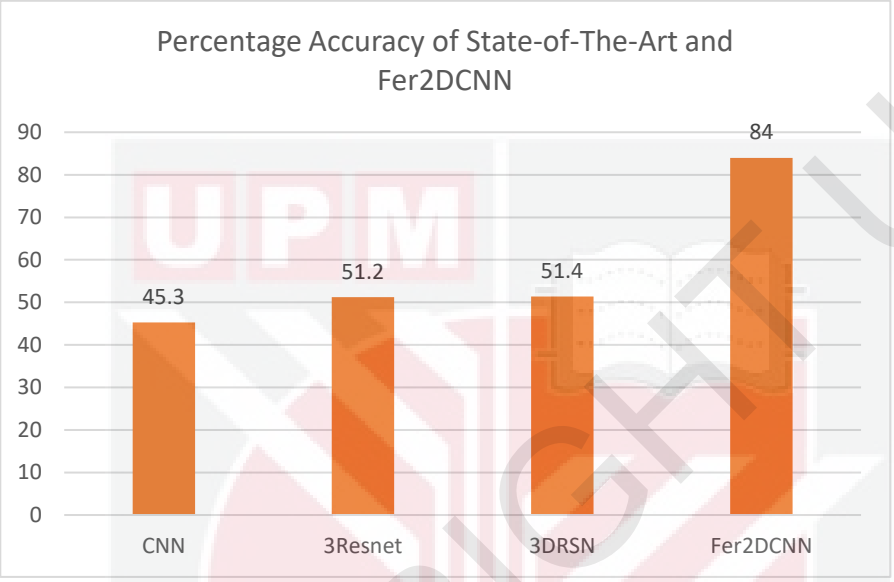

Figure 7.11 : Percentage Accuracy of State-of-The-Art and Proposed Fer2DCNN

7.5 Analysis and Conclusion

The proposed Fer2DCNN achieves significantly better performance compared to the 3DCNN model and several current state-of-the-art image and video classification models for emotion recognition (refer Table 7-5).

One of the main factors of this outstanding achievement is the emphasizing of facial expression detection using Haar Cascade Classifier which have been concluded in the development of the proposed method. The facial expressions are reliable indicators of emotions and thus serve as a strong foundation for building effective emotion recognition systems. While other cues, such as body language or speech, can provide

additional information about emotions, facial expressions often provide the most immediate and direct indication of a person's emotional state. By focusing on facial expressions, the Fer2DCNN model can quickly grasp the emotional context, enabling more accurate emotion classification.

Despite excluding the spatio-temporal element from the model development and concentrating just on the use of 2DCNN as opposed to other classification models like 3DCNN, the model was still able to outperform other models and achieved a good result. This remarkable accomplishment was made possible by taking into account the important criteria discussed below:

i. *Dataset Used Lack of Temporal Dependencies:* Emotions can often be inferred from static facial expressions captured in individual frames of a video. In this case, the temporal dynamics of facial expressions are not critical for accurately recognizing emotions in the dataset and therefore, the 2DCNN can be effective. 2DCNNs excel at capturing spatial patterns within individual frames, while ignoring temporal information.

ii. *Limited Training Data Used*: Training a 3D CNN requires a substantial amount of labelled video data to effectively learn spatiotemporal features. However, the dataset is relatively small and due to that, based on the experiment conducted earlier, the 2DCNN have shown a better performance compared to 3DCNN. One of the main factors is the 2DCNN typically require fewer parameters and are more data-efficient, making them better suited for limited training data scenarios.

iii. *Exclusion of spatiotemporal features complexity:* Emotion recognition may not heavily rely on complex spatiotemporal patterns present in videos. If emotions can be adequately represented by analyzing individual frames in isolation, a 2D CNN can be sufficient. On the other hand, if emotions are influenced by intricate spatiotemporal patterns, a 3D CNN can better capture these dynamics.

iv. *Computational Efficiency in 2DCNN:* 2DCNN is computationally less demanding than 3DCNN. Training and inference with 3DCNN can be computationally intensive due to the additional temporal dimension. In this case where low-latency performance is crucial, 2DCNN is preferred due to their lower computational requirements.

Even though this study was able to achieve outstanding outcomes with the proposed method, this study is still facing some challenges and difficulties. One of these is that when the Haar Cascade Classifier was used in the proposed method, the facial expressions that are captured from single frames appear to be very small in size. Due to this, the image resolution degrades, making several facial expressions appear similar and leading to a mismatch between the label and the image. Taking this issue into consideration, this study will continue addressing it in the next future work.

7.6 Summary of Chapter 7

This chapter focuses on the development and evaluation of the Fer2DCNN model for emotion recognition using video data from the IEMOCAP dataset. The chapter begins by addressing challenges in video-based emotion recognition such as overfitting and variability in facial expressions. To overcome these issues, the study emphasizes advanced video preprocessing, including frame extraction and facial expression detection using the Haar Cascade Classifier, which isolates facial regions for improved emotion classification.

The Fer2DCNN model processes individual video frames as images, leveraging 2D convolutional layers to extract spatial features for emotion recognition. Unlike 3DCNN models, which include spatiotemporal features, Fer2DCNN focuses solely on

spatial information, reducing overfitting risks, particularly in datasets with limited temporal dependencies.

The chapter evaluates the Fer2DCNN model through two experimental designs. The first compares Fer2DCNN with a standard 3DCNN model, revealing that Fer2DCNN achieves significantly higher accuracy compared to 3DCNN. The second experiment compares Fer2DCNN with state-of-the-art methods, including CNN and 3D CNN-based models like 3DResNet and 3DRSN, demonstrating that Fer2DCNN outperforms these methods.

In conclusion, the integration of the advanced pre-processing approach the emphasizing on facial expression detection using Haar Cascade classifier with 2DCNN enhance the performance of the proposed Fer2DCNN in emotion recognition. This combination of techniques leverages the strengths of both approaches, leading to improved accuracy in emotion recognition.

CHAPTER 8

CONCLUSION AND FUTURE WORK

8.1 Conclusion

In conclusion, this study has presented a comprehensive exploration into the field of multimodal emotion recognition. This study has explored the potential of hierarchical weighted late fusion approach as the fusion strategy in integrating audio, text, and video modalities for multimodal emotion recognition. As a result, this study has successfully proposed a great multimodal fusion model for emotion recognition known as DHWLF model. The results demonstrated that the DHWLF model outperformed both the individual models and previous state-of-the-art multimodal fusion models in terms of emotion recognition accuracy.

Apart from that, this study also has successfully proposed three more models for individual emotion recognition which are SMOTE-2DCNN model for the audio modality, TextBERT model for the text modality and Fer2DCNN model for video image modality.

8.2 Summary of Significant Contribution

The significant contribution of this study lies in the following aspects:

i. This study addresses the limitations on determining the best rules for fusing individually trained models.

For a number of reasons, the hierarchical weighted late fusion approach in DHWLF model is the best way to address the problem of fine-grained temporal and spatial information loss in late fusion approach. Firstly, the hierarchical structure allows for

the capture of both global and local dependencies within and across modalities to ensure that nuanced temporal and spatial cues are not overlooked. Through hierarchical integration of information from audio, text, and video image data, the complementary nature of these modalities can be used to improve comprehensive comprehension and representation of multimodal data thus, ensure that no fine-grained details are lost in the process.

Secondly, the weighted late fusion approach allows for the effective integration of diverse and complementary information from each modality. By assigning weights to each modality's contributions, the prioritization of the most informative features is achievable. This process preserves the fine-grained temporal and spatial information while reducing the influence of less relevant or noisy cues. With that, it ensures that the fusion process captures the most important aspects of the data while minimizing the impact of less reliable or redundant information, thereby enhancing the retention of fine-grained details.

Finally, the chosen late fusion strategy offers flexibility in the fusion process by enabling the integration of pre-trained individual emotion recognition models specialized in each modality. This approach takes advantage of the advancements in individual modality-specific models and combines their outputs at a later stage. By leveraging these specialized models, this study can benefit from each modality’s respective strengths without compromising the overall performance

ii. This study addresses data imbalance among modalities and ensure a more equitable distribution of information during fusion by incorporating data augmentation techniques.

This study has implemented various strategies to overcome the imbalance in class data availability and quality. One approach taken at the beginning of the data preparation phase involves implementing data augmentation techniques such as oversampling. During the data preparation phase, this study merged the excitement and happiness emotions to increase the amount of data as there are certain degree of resemblances inherent between both of them (refer Chapter 3). By augmenting the less represented modalities with additional samples, this study aims to rebalance the dataset and ensure a more equitable distribution of information across modalities.

Moreover, to tackle the issue of class imbalance at the individual emotion recognition model level, this study has employed techniques such as Synthetic Minority Over-sampling Technique (SMOTE) in audio emotion recognition. By synthetically generating samples for minority classes within each modality, SMOTE helps to alleviate the imbalance distribution problem thus, enhancing the robustness and fairness of the emotion recognition models across all modalities.

In addition to data augmentation techniques, another effort explored by this study is the assignment of weights to each modality based on its importance or relevance in emotion recognition. By assigning higher weights to modalities with superior data quality or greater significance in capturing emotional cues, this study sought to balance the influence of each modality during fusion. This approach ensures that no single modality dominates the fusion process. As a result, this process prevents biased representations of emotions that arise from an unequal distribution of information.

iii. This study tackles the challenge of reliance late fusion approach on prior individual model training performance by proposing the development of individual models using deep learning methods, with a focus on extracting salient features from each modality while also capturing contextual dependencies in emotion recognition tasks

This study advocate for the development of individual models using deep learning method. Specifically, the focus lies in leveraging deep learning techniques to extract salient features from each modality involving audio, text and video image data. Furthermore, the approach implemented also emphasizes on capturing contextual dependencies in emotion recognition task.

Firstly, the development for audio emotion recognition known as SMOTE-2DCNN. This model has selected MFCC along with delta and delta-delta features as optimal input for audio speech emotion recognition. MFCC captures the spectral characteristics of audio speech signals efficiently, while incorporating delta and delta-delta features helps in capturing temporal changes in the audio speech signal. By leveraging these features, the model can effectively extract salient characteristics from the audio signals.

Secondly, for the text emotion recognition model known as TextBERT model employed BERT language model to extract textual features. By utilizing BERT, salient features can be extracted by encoding the input text and generating contextualized word embeddings, which capture rich information about the relationships between words. These embeddings enable the identification of key features, such as entities, sentiment, or specific topics, within the text. In addition, BERT's attention mechanism allows it to assign varying levels of importance to

different parts of the input, ensuring that salient features are accurately identified and represented.

Furthermore, this study introduced Fer2DCNN for video image emotion recognition. In this model's approach, the implementation involved a Haar cascade classifier which enable an accurate detection and extraction of the facial region. By isolating the face and excluding other irrelevant parts, the mitigation of overfitting risk and enhancement of specificity in the features extracted from the visual modality were achieved. The Fer2DCNN approach provides a robust and efficient means of extracting visual cues related to human emotions.

Apart from that, in term of contextual learning, the introduction of TextBERT model possesses the capability to capture and comprehend the intricate relationships and dependencies within the provided text. By leveraging TextBERT's contextual understanding, the accuracy and performance of emotion recognition tasks were improved through effective modeling and capturing of the long-term dependencies inherent in contextual information. This approach does not only provide a more robust and accurate solution to emotion recognition, but it also offers a theoretical foundation and practical application for incorporating BERT language model into deep learning models for similar tasks.

8.3 Limitation and Future Direction

Although this study has successfully addressed all the significant problems and achieved all the objectives outlined in Chapter 1, there are a few limitations to consider.

One limitation is the reliance on predefined weights for multimodal fusion. While the weighted late fusion approach allows for flexibility, determining the optimal weights can be subjective and challenging. Future studies could explore approaches that automatically learn the weights based on the data and task at hand, potentially leveraging deep learning techniques to adaptively determine the modality importance levels.

Furthermore, this proposed method primarily focuses on fusing audio, text, and visual modalities. However, as multimodal emotion recognition analysis continues to evolve, there may be other modalities or data sources that could provide valuable insights. Exploring the fusion of additional modalities, such as sensor data, social media interactions, or physiological signals, could lead to more comprehensive and informative representations of multimodal data.

Lastly, the evaluation of the proposed method should be extended to various real-world applications and datasets. Further studies are needed to validate its performance across different domains, tasks, and datasets. This includes considering the robustness of the approach under various noise and domain shift scenarios.

By addressing these limitations and exploring the proposed future directions, researchers can further advance multimodal fusion strategies and contribute to the development of more sophisticated and practical solutions for emotion recognition analysis which involves various modalities of data.

REFERENCES

Abayomi-Alli, O. O., Damaševičius, R., Qazi, A., Adedoyin-Olowe, M., & Misra, S. (2022). Data augmentation and deep learning methods in sound classification: A systematic review. *Electronics*, *11*(22), 3795.

Agrawal, A., & Mittal, N. (2020). Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy. *The Visual Computer*, *36*(2), 405–412.

Akçay, M. B., & Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, *116*, 56–76.

Akinduyite, O. C., Badeji-Ajisafe, B., Oguntimilehin, A., Obamiyi, S. E., Ajibade, A. O., & Fikayomi, A. O. (2024). Facial Emotion Recognition Using Convolutional Neural Network in a Learning Environment. *2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG)*, 1–8. https://doi.org/10.1109/SEB4SDG60871.2024.10630178

Ancilin, J., & Milton, A. (2021). Improved speech emotion recognition with Mel frequency magnitude coefficient. *Applied Acoustics*, *179*, 108046.

Anvarjon, T., Mustaqeem, & Kwon, S. (2020). Deep-net: A lightweight CNN-based speech emotion recognition system using deep frequency features. *Sensors*, *20*(18), 5212.

Aridas, C. K., Karlos, S., Kanas, V. G., Fazakis, N., & Kotsiantis, S. B. (2019). Uncertainty based under-sampling for learning naive bayes classifiers under imbalanced data sets. *IEEE Access*, *8*, 2122–2133.

Bharti, S. K., & Babu, K. S. (2018). Sarcasm as a contradiction between a tweet and its temporal facts: a pattern-based approach. *International Journal on Natural Language Computing (IJNLC) Vol*, *7*.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, *42*, 335–359.

Cambria, E., Poria, S., Gelbukh, A., & Thelwall, M. (2017). Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, *32*(6), 74–80.

Chao, L., Tao, J., Yang, M., Li, Y., & Wen, Z. (2015). Long short term memory recurrent neural network based multimodal dimensional emotion recognition. *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, 65–72.

Chatterjee, M., Zion, D. J., Deroche, M. L., Burianek, B. A., Limb, C. J., Goren, A. P., Kulkarni, A. M., & Christensen, J. A. (2015). Voice emotion recognition

by cochlear-implanted children and their normally-hearing peers. *Hearing Research*, *322*, 151–162.

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002a). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002b). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, L.-W., & Rudnicky, A. (2023). Exploring Wav2vec 2.0 Fine Tuning for Improved Speech Emotion Recognition. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10095036

Chennuru, V. K., & Timmapareddy, S. R. (2022). Simulated annealing based undersampling (SAUS): A hybrid multi-objective optimization method to tackle class imbalance. *Applied Intelligence*, *52*(2), 2092–2110.

Chen, S., & Jin, Q. (2015). Multi-modal dimensional emotion recognition using recurrent neural networks. *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, 49–56.

Chinchanikar, N. A. (2019). Facial Expression Recognition Using Deep Learning: A Review. *International Research Journal of Engineering and Technology (IRJET)*, *6*, 3274–3281.

Choi, D. Y., & Song, B. C. (2020). Semi-supervised learning for continuous emotion recognition based on metric learning. *IEEE Access*, *8*, 113443–113455.

Choi, H.-S., Jung, D., Kim, S., & Yoon, S. (2021). Imbalanced data classification via cooperative interaction between classifier and generator. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(8), 3343–3356.

Chopade, C. R. (2015). Text based emotion recognition: A survey. *International Journal of Science and Research*, *4*(6), 409–414.

Cimtay, Y., Ekmekcioglu, E., & Caglar-Ozhan, S. (2020). Cross-subject multimodal emotion recognition based on hybrid fusion. *IEEE Access*, *8*, 168865–168878.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, *18*(1), 32–80.

Dablain, D., Krawczyk, B., & Chawla, N. V. (2022). DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*.

Datcu, D., & Rothkrantz, L. J. M. (2015). Semantic audiovisual data fusion for automatic emotion recognition. *Emotion Recognition: A Pattern Analysis Approach*, 411–435.

Deng, J., & Ren, F. (2021). A survey of textual emotion recognition and its challenges. *IEEE Transactions on Affective Computing*.

Devi, D., Biswas, S. K., & Purkayastha, B. (2022). Correlation-based oversampling aided cost sensitive ensemble learning technique for treatment of class imbalance. *Journal of Experimental & Theoretical Artificial Intelligence*, *34*(1), 143–174.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805*.

Dhall, A., Goecke, R., Lucey, S., & Gedeon, T. (2012). Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimedia*, *19*(3), 34.

Dinakaran, K., & Ashokkrishna, E. M. (2020). Efficient regional multi feature similarity measure based emotion detection system in web portal using artificial neural network. *Microprocessors and Microsystems*, *77*, 103112.

Duc, B., Bigün, E. S., Bigün, J., Maître, G., & Fischer, S. (1997). Fusion of audio and video information for multi modal person authentication. *Pattern Recognition Letters*, *18*(9), 835–843.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, *6*(3–4), 169–200.

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.

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological Bulletin*, *128*(2), 203.

Elreedy, D., Atiya, A. F., & Kamalov, F. (2023). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Machine Learning*, 1–21.

Elyan, E., Moreno-Garcia, C. F., & Jayne, C. (2021). CDSMOTE: class decomposition and synthetic minority class oversampling technique for imbalanced-data classification. *Neural Computing and Applications*, *33*, 2839–2851.

Fan, Y., Lu, X., Li, D., & Liu, Y. (2016). Video-based emotion recognition using CNN-RNN and C3D hybrid networks. *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 445–450.

Ganaie, M. A., Tanveer, M., & Initiative, A. D. N. (2021). Fuzzy least squares projection twin support vector machines for class imbalance learning. *Applied Soft Computing*, *113*, 107933.

Ganesh, P., Chen, Y., Lou, X., Khan, M. A., Yang, Y., Sajjad, H., Nakov, P., Chen, D., & Winslett, M. (2021). Compressing large-scale transformer-based models: A case study on bert. *Transactions of the Association for Computational Linguistics*, *9*, 1061–1080.

Ghazouani, H. (2021). A genetic programming-based feature selection and fusion for facial expression recognition. *Applied Soft Computing*, *103*, 107173.

Ghosal, D., Majumder, N., Poria, S., Chhaya, N., & Gelbukh, A. (2019). Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *ArXiv Preprint ArXiv:1908.11540*.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., & Lee, D.-H. (2013). Challenges in representation learning: A report on three machine learning contests. *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*, 117–124.

Gupta, A., & Batra, S. (2023). Real-Time Emotion Recognition using Deep Facial Expression Analysis on Mobile Devices. *Research Journal of Computer Systems and Engineering*, *4*(2), 89–102.

Gupta, R., Khomami Abadi, M., Cárdenes Cabré, J. A., Morreale, F., Falk, T. H., & Sebe, N. (2016). A quality adaptive multimodal affect recognition system for user-centric multimedia indexing. *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, 317–320.

Gu, Y., Yang, K., Fu, S., Chen, S., Li, X., & Marsic, I. (2018). Multimodal Affective Analysis Using Hierarchical Attention Strategy with Word-Level Alignment. *Proceedings of the Conference. Association for Computational Linguistics. Meeting*, *2018*, 2225–2235.

Guzmán-Ponce, A., Sánchez, J. S., Valdovinos, R. M., & Marcial-Romero, J. R. (2021). DBIG-US: A two-stage under-sampling algorithm to face the class imbalance problem. *Expert Systems with Applications*, *168*, 114301.

Hazarika, D., Gorantla, S., Poria, S., & Zimmermann, R. (2018). Self-attentive feature-level fusion for multimodal emotion detection. *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 196–201.

Hazarika, D., Poria, S., Zadeh, A., Cambria, E., Morency, L.-P., & Zimmermann, R. (2018). Conversational memory network for emotion recognition in dyadic dialogue videos. *Proceedings of the Conference. Association for Computational Linguistics. North American Chapter. Meeting*, *2018*, 2122.

Ho, N.-H., Yang, H.-J., Kim, S.-H., & Lee, G. (2020). Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network. *IEEE Access*, *8*, 61672–61686.

Huang, Y., Yang, J., Liao, P., & Pan, J. (2017a). Fusion of facial expressions and EEG for multimodal emotion recognition. *Computational Intelligence and Neuroscience*, *2017*.

Huang, Y., Yang, J., Liao, P., & Pan, J. (2017b). Fusion of facial expressions and EEG for multimodal emotion recognition. *Computational Intelligence and Neuroscience*, *2017*.

Huang, Z., Dong, M., Mao, Q., & Zhan, Y. (2014). Speech emotion recognition using CNN. *Proceedings of the 22nd ACM International Conference on Multimedia*, 801–804.

Jain, D. K., Shamsolmoali, P., & Sehdev, P. (2019). Extended deep neural network for facial emotion recognition. *Pattern Recognition Letters*, *120*, 69–74.

James, W. (1884). On some omissions of introspective psychology. *Mind*, *9*(33), 1–26.

Jiang, Y., Li, W., Hossain, M. S., Chen, M., Alelaiwi, A., & Al-Hammadi, M. (2020). A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition. *Information Fusion*, *53*, 209–221.

Jiao, W., Yang, H., King, I., & Lyu, M. R. (2019). HiGRU: hierarchical gated recurrent units for utterance-level emotion recognition. *ArXiv Preprint ArXiv:1904.04446*.

Jung, H., Lee, S., Yim, J., Park, S., & Kim, J. (2015). Joint fine-tuning in deep neural networks for facial expression recognition. *Proceedings of the IEEE International Conference on Computer Vision*, 2983–2991.

Kaiser, J. F. (1990). On a simple algorithm to calculate the'energy'of a signal. *International Conference on Acoustics, Speech, and Signal Processing*, 381–384.

Karpagavalli, S., & Chandra, E. (2016). A review on automatic speech recognition architecture and approaches. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, *9*(4), 393–404.

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.

Kim, D. H., Baddar, W. J., Jang, J., & Ro, Y. M. (2017). Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. *IEEE Transactions on Affective Computing*, *10*(2), 223–236.

Kim, T., Lee, J., & Nam, J. (2019). Comparison and analysis of SampleCNN architectures for audio classification. *IEEE Journal of Selected Topics in Signal Processing*, *13*(2), 285–297.

Kohler, C. G., Turner, T. H., Gur, R. E., & Gur, R. C. (2004). Recognition of facial emotions in neuropsychiatric disorders. *CNS Spectrums*, *9*(4), 267–274.

Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: a review. *International Journal of Speech Technology*, *15*, 99–117.

Koziarski, M., Bellinger, C., & Woźniak, M. (2021). RB-CCR: Radial-Based Combined Cleaning and Resampling algorithm for imbalanced data classification. *Machine Learning*, *110*, 3059–3093.

Kuncheva, L. I. (2014). *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons.

Lee, C. M., & Narayanan, S. S. (2005). Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, *13*(2), 293–303.

Liang, D., Liang, H., Yu, Z., & Zhang, Y. (2020). Deep convolutional BiLSTM fusion network for facial expression recognition. *The Visual Computer*, *36*, 499–508.

Li, J., Ji, D., Li, F., Zhang, M., & Liu, Y. (2020). Hitrans: A transformer-based context-and speaker-sensitive model for emotion detection in conversations. *Proceedings of the 28th International Conference on Computational Linguistics*, 4190–4200.

Li, J., Wang, S., Chao, Y., Liu, X., & Meng, H. (2022). Context-aware Multimodal Fusion for Emotion Recognition. *INTERSPEECH*, 2013–2017.

Li, Q., Wu, C., Wang, Z., & Zheng, K. (2020). Hierarchical transformer network for utterance-level emotion recognition. *Applied Sciences*, *10*(13), 4447.

Li, S., & Tang, H. (2024). Multimodal Alignment and Fusion: A Survey. *ArXiv Preprint ArXiv:2411.17040*.

Li, Y., Zeng, J., Shan, S., & Chen, X. (2018). Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Transactions on Image Processing*, *28*(5), 2439–2450.

Lopes, A. T., De Aguiar, E., De Souza, A. F., & Oliveira-Santos, T. (2017). Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. *Pattern Recognition*, *61*, 610–628.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. *2010 Ieee Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, 94–101.

Lyons, M. J. (2021). " Excavating AI" Re-excavated: Debunking a Fallacious Account of the JAFFE Dataset. *ArXiv Preprint ArXiv:2107.13998*.

Majumder, N., Hazarika, D., Gelbukh, A., Cambria, E., & Poria, S. (2018). Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-Based Systems*, *161*, 124–133.

Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., & Cambria, E. (2019). Dialoguernn: An attentive rnn for emotion detection in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 6818–6825.

Malygina, M., Artemyev, M., Belyaev, A., & Perepelkina, O. (2019). *Overview of the Advancements in Automatic Emotion Recognition: Comparative Performance of Commercial Algorithms*.

Mansour, A., & Lachiri, Z. (2017). SVM based emotional speaker recognition using MFCC-SDC features. *International Journal of Advanced Computer Science and Applications*, *8*(4).

Mao, Q., Dong, M., Huang, Z., & Zhan, Y. (2014). Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Transactions on Multimedia*, *16*(8), 2203–2213.

Mayabadi, S., & Saadatfar, H. (2022). Two density-based sampling approaches for imbalanced and overlapping data. *Knowledge-Based Systems*, *241*, 108217.

Ma, Y., Hao, Y., Chen, M., Chen, J., Lu, P., & Košir, A. (2019). Audio-visual emotion fusion (AVEF): A deep efficient weighted approach. *Information Fusion*, *46*, 184–192.

McCann, B., Bradbury, J., Xiong, C., & Socher, R. (2017). Learned in translation: Contextualized word vectors. *Advances in Neural Information Processing Systems*, *30*.

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, *14*, 261–292.

Meng, T., Shou, Y., Ai, W., Yin, N., & Li, K. (2023). Deep imbalanced learning for multimodal emotion recognition in conversations. *ArXiv Preprint ArXiv:2312.06337*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *ArXiv Preprint ArXiv:1301.3781*.

Moin, A., Aadil, F., Ali, Z., & Kang, D. (2023). Emotion recognition framework using multiple modalities for an effective human–computer interaction. *The Journal of Supercomputing*, *79*(8), 9320–9349.

Nandi, A., & Xhafa, F. (2022). A federated learning method for real-time emotion state classification from multi-modal streaming. *Methods*, *204*, 340–347.

Nguyen, D., Nguyen, K., Sridharan, S., Dean, D., & Fookes, C. (2018). Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition. *Computer Vision and Image Understanding*, *174*, 33–42.

Nwe, T. L., Foo, S. W., & De Silva, L. C. (2003a). Detection of stress and emotion in speech using traditional and FFT based log energy features. *Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint*, *3*, 1619–1623.

Nwe, T. L., Foo, S. W., & De Silva, L. C. (2003b). Speech emotion recognition using hidden Markov models. *Speech Communication*, *41*(4), 603–623.

Pandeya, Y. R., & Lee, J. (2021). Deep learning-based late fusion of multimodal information for emotion classification of music video. *Multimedia Tools and Applications*, *80*, 2887–2905.

Panksepp, J. (2004). *Affective neuroscience: The foundations of human and animal emotions*. Oxford university press.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (1802). Deep contextualized word representations. CoRR abs/1802.05365 (2018). *ArXiv Preprint ArXiv:1802.05365*.

Pfeuffer, A., & Dietmayer, K. (2018). Optimal sensor data fusion architecture for object detection in adverse weather conditions. *2018 21st International Conference on Information Fusion (FUSION)*, 1–8.

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, *37*, 98–125.

Poria, S., Cambria, E., Howard, N., Huang, G.-B., & Hussain, A. (2016). Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, *174*, 50–59.

Poria, S., Chaturvedi, I., Cambria, E., & Hussain, A. (2016). Convolutional MKL based multimodal emotion recognition and sentiment analysis. *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 439–448.

Poria, S., Majumder, N., Hazarika, D., Cambria, E., Gelbukh, A., & Hussain, A. (2018). Multimodal sentiment analysis: Addressing key issues and setting up the baselines. *IEEE Intelligent Systems*, *33*(6), 17–25.

Ramachandram, D., & Taylor, G. W. (2017). Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, *34*(6), 96–108.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161.

Sahay, S., Kumar, S. H., Xia, R., Huang, J., & Nachman, L. (2018). Multimodal relational tensor network for sentiment and emotion classification. *ArXiv Preprint ArXiv:1806.02923*.

Santoso, J., Yamada, T., Ishizuka, K., Hashimoto, T., & Makino, S. (2022). Speech Emotion Recognition Based on Self-Attention Weight Correction for Acoustic and Text Features. *IEEE Access*, *10*, 115732–115743.

Satt, A., Rozenberg, S., & Hoory, R. (2017). Efficient emotion recognition from speech using deep learning on spectrograms. *Interspeech*, 1089–1093.

Schroder, M., & Cowie, R. (2006). Issues in emotion-oriented computing-towards a shared understanding. *Workshop on Emotion and Computing*.

Schuller, B. W. (2018). Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, *61*(5), 90–99.

Sebe, N. (2005). *Machine learning in computer vision* (Vol. 29). Springer Science & Business Media.

Shankar, S., Thompson, L., & Fiterau, M. (2022). Progressive Fusion for Multimodal Integration. *ArXiv Preprint ArXiv:2209.00302*.

Shaver, P. , S. J. , K. D. , & O. C. (1987). Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, *52*(6), 1061–1086.

Singh, M., & Fang, Y. (2020). Emotion recognition in audio and video using deep neural networks. *ArXiv Preprint ArXiv:2006.08129*.

Singh, Y. B., & Goel, S. (2022). A systematic literature review of speech emotion recognition approaches. *Neurocomputing*, *492*, 245–263.

Sreevidya, P., Aravinth, J., & Samiappan, S. (2024). Intermediate layer attention mechanism for multimodal fusion in personality and affect computing. *IEEE Access*.

Suen, C. Y., & Lam, L. (2000). Multiple classifier combination methodologies for different output levels. *International Workshop on Multiple Classifier Systems*, 52–66.

Sun, B., Li, L., Zhou, G., Wu, X., He, J., Yu, L., Li, D., & Wei, Q. (2015). Combining multimodal features within a fusion network for emotion recognition in the wild. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 497–502.

Sun, D., He, Y., & Han, J. (2023). Using Auxiliary Tasks In Multimodal Fusion of Wav2vec 2.0 And Bert for Multimodal Emotion Recognition. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096586

Swain, M., Routray, A., & Kabisatpathy, P. (2018). Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, *21*, 93–120.

Teager, H. (1980). Some observations on oral air flow during phonation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(5), 599–601.

Thejas, G. S., Hariprasad, Y., Iyengar, S. S., Sunitha, N. R., Badrinath, P., & Chennupati, S. (2022). An extension of Synthetic Minority Oversampling Technique based on Kalman filter for imbalanced datasets. *Machine Learning with Applications*, *8*, 100267.

Tripathi, S., Tripathi, S., & Beigi, H. (2019). MULTI-MODAL EMOTION RECOGNITION ON IEMOCAP WITH NEURAL NETWORKS. *ArXiv Preprint ArXiv:1804.05788*.

Tsalera, E., Papadakis, A., & Samarakou, M. (2021). Comparison of pre-trained CNNs for audio classification using transfer learning. *Journal of Sensor and Actuator Networks*, *10*(4), 72.

Tsanousa, A., Meditskos, G., Vrochidis, S., & Kompatsiaris, I. (2019). A weighted late fusion framework for recognizing human activity from wearable sensors. *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, 1–8.

Tzirakis, P., Chen, J., Zafeiriou, S., & Schuller, B. (2021). End-to-end multimodal affect recognition in real-world environments. *Information Fusion*, *68*, 46–53.

Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., & Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, *11*(8), 1301–1309.

Tzirakis, P., Zhang, J., & Schuller, B. W. (2018). End-to-end speech emotion recognition using deep neural networks. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5089–5093.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, *48*(9), 1162–1181.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, *1*, I–I.

Vomberg, L. H. (2019). *Mechanisms of pitch discrimination in musicians and non-musicians*. University of Lethbridge (Canada).

Wagner, J., Andre, E., Lingenfelser, F., & Kim, J. (2011). Exploring fusion methods for multimodal emotion recognition with missing data. *IEEE Transactions on Affective Computing*, *2*(4), 206–218.

Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., & Ambikairajah, E. (2021a). A comprehensive review of speech emotion recognition systems. *IEEE Access*, *9*, 47795–47814.

Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., & Ambikairajah, E. (2021b). A comprehensive review of speech emotion recognition systems. *IEEE Access*, *9*, 47795–47814.

Wu, C.-H., & Liang, W.-B. (2010). Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transactions on Affective Computing*, *2*(1), 10–21.

Xi, D., Tang, L., Chen, R., & Xu, W. (2023). A multimodal time-series method for gifting prediction in live streaming platforms. *Information Processing & Management*, *60*(3), 103254.

Xie, X., Liu, H., Zeng, S., Lin, L., & Li, W. (2021). A novel progressively undersampling method based on the density peaks sequence for imbalanced data. *Knowledge-Based Systems*, *213*, 106689.

Xu, H., Zhang, H., Han, K., Wang, Y., Peng, Y., & Li, X. (2019). Learning alignment for multimodal emotion recognition from speech. *ArXiv Preprint ArXiv:1909.05645*.

Xu, X., Wang, T., Yang, Y., Zuo, L., Shen, F., & Shen, H. T. (2020). Cross-modal attention with semantic consistence for image–text matching. *IEEE Transactions on Neural Networks and Learning Systems*, *31*(12), 5412–5425.

Yang, H., Ciftci, U., & Yin, L. (2018). Facial expression recognition by de-expression residue learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2168–2177.

Yolcu, G., Oztel, I., Kazan, S., Oz, C., Palaniappan, K., Lever, T. E., & Bunyak, F. (2019). Facial expression recognition for monitoring neurological disorders based on convolutional neural network. *Multimedia Tools and Applications*, *78*, 31581–31603.

Yu, Z., Liu, G., Liu, Q., & Deng, J. (2018). Spatio-temporal convolutional features with nested LSTM for facial expression recognition. *Neurocomputing*, *317*, 50–57.

Zhalehpour, S., Onder, O., Akhtar, Z., & Erdem, C. E. (2016). BAUM-1: A spontaneous audio-visual face database of affective and mental states. *IEEE Transactions on Affective Computing*, *8*(3), 300–313.

Zhang, D., Wu, L., Sun, C., Li, S., Zhu, Q., & Zhou, G. (2019). Modeling both Context-and Speaker-Sensitive Dependence for Emotion Detection in Multi-speaker Conversations. *IJCAI*, 5415–5421.

Zhang, J., Xing, L., Tan, Z., Wang, H., & Wang, K. (2022). Multi-head attention fusion networks for multi-modal speech emotion recognition. *Computers & Industrial Engineering*, *168*, 108078.

Zhang, S., Yang, Y., Chen, C., Zhang, X., Leng, Q., & Zhao, X. (2024). Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects. *Expert Systems with Applications*, *237*, 121692.

Zhang, S., Zhang, S., Huang, T., Gao, W., & Tian, Q. (2017). Learning affective features with a hybrid deep model for audio–visual emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *28*(10), 3030–3043.

Zhang, W., Zhao, D., Chai, Z., Yang, L. T., Liu, X., Gong, F., & Yang, S. (2017). Deep learning and SVM-based emotion recognition from Chinese speech for smart affective services. *Software: Practice and Experience*, *47*(8), 1127–1138.

Zhao, G., Zhang, Y., & Chu, J. (2023). Emo-BERT: A Multi-Modal Teacher Speech Emotion Recognition Method. *2023 5th International Workshop on Artificial Intelligence and Education (WAIE)*, 80–84. https://doi.org/10.1109/WAIE60568.2023.00022

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.

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.

Zheng, M., Li, T., Sun, L., Wang, T., Jie, B., Yang, W., Tang, M., & Lv, C. (2021). An automatic sampling ratio detection method based on genetic algorithm for imbalanced data classification. *Knowledge-Based Systems*, *216*, 106800.

Zhou, F., Kong, S., Fowlkes, C. C., Chen, T., & Lei, B. (2020). Fine-grained facial expression analysis using dimensional emotion model. *Neurocomputing*, *392*, 38–49.

Ziemer, K. S., & Korkmaz, G. (2017). Using text to predict psychological and physical health: A comparison of human raters and computerized text analysis. *Computers in Human Behavior*, *76*, 122–127.

BIODATA OF STUDENT

Nurul Nadhrah binti Kamaruzaman was born in November 1989 in Kuala Lumpur. In 2012, she successfully obtained a Bachelor's degree in Computer Science from the esteemed University of Malaya (UM), where she specialized in software engineering.

Nurul Nadhrah pursued further education and earned a Master's degree in Computer Science from the University of Malaya in 2016. During her master's program, she focused on the domain of human-computer interaction, specializing in the development of serious games to assist students with autism in their learning process. This research highlighted her commitment to leveraging technology for the betterment of individuals with special needs.

In 2018, Nurul Nadhrah continued her academic journey at Universiti Putra Malaysia (UPM) to pursue a Ph.D degree in intelligent computing. Her studies at UPM expanded her knowledge and expertise, particularly in the realm of multimodal fusion models for emotion recognition, with a specific focus on integrating audio, text, and video. She showcased a keen interest in deep learning, which further demonstrated her determination to explore cutting-edge technologies and their applications.

While excelling academically, Nurul Nadhrah also gained valuable practical experience in the field. She worked as a tutor at Pusat Asasi Sains Universiti Malaya in 2018 until 2019. In 2019, she took on the role of a part-time lecturer at City University, where she was entrusted with teaching subjects such as Discrete Mathematics, Artificial Intelligence, and Mobile Computing.

Additionally, Nurul Nadhrah served as a lab demonstrator in the Faculty of Computer Science at Universiti Putra Malaysia since 2019. She was responsible for assisting students in subjects such as Programming 1 and Artificial Intelligence, providing them with practical insights and guidance in their learning process.

LIST OF PUBLICATIONS

Article in Journal

Kamaruzaman, N. N., Husin, N. A., Mustapha, N., Yaakob, R., Ejaz, M. M., Hassan, R., ... & Khare, A. (2024). SMOTE-2DCNN For Enhancing Speech Emotion Recognition. *Journal of Theoretical and Applied Information Technology*, *102*(13).

Conferences

Kamaruzaman, Nurul. (2019). Towards a Multimodal Analysis to Predict Mental Illness in Twitter Platform. International Journal of Advanced Trends in Computer Science and Engineering. 126-130. 10.30534/ijatcse/2019/1981.42019.