Audio-Visual Feature Fusion through Transformers for Automated Depression Screening in Social Media Content

by

Md Rezwanul Haque

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2026

© Md Rezwanul Haque 2026

Author's Declaration

This thesis consists of materials all of which I authored and co-authored: see the *Statement of Contributions* included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

The research presented in this thesis has resulted in the following publications:

1. **Md Rezwanul Haque**, Md. Milon Islam, S. M. Taslim Uddin Raju, Hamdi Altaheri, Lobna Nassar and Fakhri Karray, "MDD-Net: Multimodal Depression Detection through Mutual Transformer," *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Vienna, Austria, 2025, pp. 6592-6599, doi: 10.1109/SMC58881.2025.11342588.

2. **Md Rezwanul Haque**, Md. Milon Islam, S. M. Taslim Uddin Raju, Hamdi Altaheri, Lobna Nassar and Fakhri Karray, "MMFformer: Multimodal Fusion Transformer Network for Depression Detection," *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Vienna, Austria, 2025, pp. 5912-5919, doi: 10.1109/SMC58881.2025.11342694.

In both papers, I took the lead on the conceptualization, methodology development, investigation processes, validation, and the drafting of the original manuscript. Md. Milon Islam provided significant contributions to the development and validation of the methodology, and the writing of the original draft.

S. M. Taslim Uddin Raju was primarily responsible for enhancing and refining the Introduction section of both papers. Hamdi Altaheri contributed to the writing through review and editing support.

Across all papers, Prof. Fakhri Karray provided thorough reviews, detailed revision, invaluable guidance, and suggestions. Meanwhile, Dr. Lobna Nassar further enhanced the work with her review, revision, guidance, and further recommendations.

Chapters 1 to 5 of this thesis incorporate content from both papers.

Abstract

Depression has become a critical public health concern, with the World Health Organization reporting that over 280 million people worldwide are affected by it. The rapid growth of social media, particularly video blogs, has drawn research attention toward analyzing user-generated audiovisual content for signs of depression. These videos capture natural facial expressions, voice characteristics, and speech patterns that may reveal more about a person's emotional state than verbal self-reports alone. However, extracting useful features from such noisy, unstructured data and combining audio and visual information in a way that preserves their complementary nature remain open problems in this domain.

The thesis is organized into two main contributions. In the first part, we propose MDD-Net, a multimodal depression detection network that uses a mutual transformer to fuse acoustic and visual features. The acoustic branch employs a global self-attention network to process 25 low-level descriptors including loudness, Mel-Frequency Cepstral Coefficients, and spectral flux, capturing both content-based and positional relationships. The visual branch applies hierarchical multi-head self-attention on 68 facial landmarks extracted from each video frame. The mutual transformer then operates bidirectionally: audio queries attend to visual keys and values, and visual queries attend to audio keys and values. We also design a composite loss function that combines binary cross-entropy, focal loss, and L2 regularization to handle the noisy labels and class imbalance that are common in social media datasets.

In the second part, we introduce MMFformer, a multimodal fusion transformer network that takes a different approach to the same problem. For video, a pre-trained vision transformer augmented with residual connections extracts high-level spatial patterns from facial data. For audio, a transformer encoder built on the audio spectrogram transformer paradigm models temporal dynamics in speech signals through patch and positional embeddings. On the fusion side, we propose and compare three distinct strategies: late transformer fusion, intermediate transformer fusion, and intermediate attention fusion, each operating at a different level of the processing pipeline.

We evaluate both architectures on the D-Vlog dataset, a publicly available collection of 961 YouTube vlogs from 816 individuals annotated for depression. MMFformer is additionally tested on the LMVD dataset, a larger corpus of 1,823 vlogs collected from four different social media platforms. MDD-Net reaches an F1-Score of 77.07% on D-Vlog, which is an improvement ranging from 1.82% to 17.37% over previously reported methods. MMFformer achieves 90.92% on D-Vlog and 90.48% on LMVD, surpassing the best existing results by 13.92% and 7.74% respectively. Cross-corpus validation between D-Vlog and LMVD further confirms that the developed architectures generalize across different platforms and populations.

Acknowledgements

In the name of Allah, Most Gracious, Most Merciful

First and foremost, I wish to convey my profound appreciation to Allah (SWT), the Almighty, for bestowing upon me the strength, patience, and perseverance required to achieve this endeavor. He has granted me opportunity, knowledge, and the capacity to tackle challenges, for which I am sincerely grateful.

I would like to express my deep gratitude to my MASc supervisors, Prof. Fakhri Karray and Prof. Pin-Han Ho, for their continuous help and support throughout my graduate journey. In particular, Prof. Fakhri Karray's broad research vision and encouragement to explore problems across multiple domains had a significant effect on my academic growth. I appreciated the freedom he offered me to pursue my research interests and the opportunities he provided to publish at top-tier venues such as IEEE SMC, ICCV, and IJCNN while working on my thesis. I am deeply grateful for his guidance, and I look forward to continuing this journey under his supervision as I begin my doctoral studies.

I would like to thank my committee members, Prof. Zhou Wang and Prof. Kumaraswamy Ponnambalam, for their valuable feedback and their time. I was extremely fortunate to have such an intellectual committee with me along the path.

I am also very grateful to Md. Milon Islam for his continuous advice and mentorship throughout my research career. May Allah reward him and his family abundantly in this world and the hereafter.

Additionally, I want to express my gratitude to my dear friends and colleagues from back home, whose encouragement, constructive criticisms, and helpful comments helped me achieve my goals. I would also like to thank the fellow members at the Centre for Pattern Analysis and Machine Intelligence (CPAMI), especially S M Taslim Uddin Raju, Hamdi Altaheri, and Dr. Lobna Nassar, for their help, support, and insightful discussions. I am proud to be a part of such an outstanding community.

Finally, I am deeply thankful to my mother, father, sisters, brother, and grandparents who were always by my side and helped me through the ups and downs of starting a new life in a new country. I am very blessed to have you all.

Dedication

This is dedicated to the ones I love, especially to my mother and father.

Table of Contents

Author's Declaration ii

Statement of Contributions iii

Abstract iv

Acknowledgements v

Dedication vi

List of Figures x

List of Tables xii

List of Abbreviations xiii

1 Introduction 1

1.1 Problem Definition 1

1.2 Motivation 3

1.3 Scope of the Work 3

1.4 Contributions 4

1.5 Thesis Organization 5

2 Literature Review **7**

2.1 Depression Detection Using Deep Learning 7

2.2 Advanced Multimodal Architectures for Depression Detection 10

2.3 Transformer-Based Approaches for Depression Detection 11

2.4 Summary and Research Gaps 12

3 Proposed Methodology **14**

3.1 MDD-Net: Multimodal Depression Detection through Mutual Transformer . . . 14

3.1.1 Problem Formulation 15

3.1.2 Acoustic Feature Extraction Module 16

3.1.3 Visual Feature Extraction Module 17

3.1.4 Mutual Transformer 19

3.1.5 Depression Detection Layer 20

3.1.6 Loss Function 21

3.2 MMFformer: Multimodal Fusion Transformer Network for Depression Detection 21

3.2.1 Video Feature Extraction 23

3.2.2 Audio Feature Extraction 25

3.2.3 Fusion Module 26

3.2.3.1 Late Transformer Fusion 26

3.2.3.2 Intermediate Transformer Fusion 27

3.2.3.3 Intermediate Attention Fusion 27

3.3 Summary 28

4 Experiments and Results **29**

4.1 Datasets 29

4.1.1 D-Vlog Dataset 29

4.1.2 LMVD Dataset 30

4.1.3 Comparison of Dataset Characteristics 30

4.2 Experimental Setup . . . 31
4.2.1 MDD-Net Implementation Details . . . 31
4.2.2 MMFformer Implementation Details . . . 32
4.2.3 Comparison of Training Configurations . . . 32
4.2.4 Evaluation Metrics . . . 33
4.3 MDD-Net Results . . . 33
4.3.1 Performance Analysis . . . 33
4.3.2 Comparison with the State-of-the-Art . . . 34
4.3.3 Ablation Study . . . 35
4.3.4 Visual Results . . . 36
4.4 MMFformer Results . . . 37
4.4.1 Depression Detection Results . . . 38
4.4.2 Comparison with the State-of-the-Art . . . 39
4.4.3 Ablation Study . . . 40
4.4.4 Cross-Corpus Validation . . . 41
4.5 Summary . . . 42

5 Conclusion and Future Work **43**
5.1 Summary of Contributions . . . 43
5.2 Limitations . . . 44
5.3 Future Work . . . 45

References **46**

APPENDICES **53**

A Architecture Configurations and Model Comparison **54**
A.1 Architecture Configuration of MDD-Net . . . 55
A.2 Architecture Configuration of MMFformer . . . 56
A.3 Comparative Summary of Model Configurations . . . 57

List of Figures

3.1 The overall architecture of the proposed MDD-Net system: a multimodal fusion technique for audio-visual feature fusion using mutual transformers for depression detection. 15

3.2 Acoustic feature extraction module: the module processes input audio features using 1×1 convolutions, applies both content and positional attention mechanisms, and performs batch normalization before fusing the features for the final representations. 17

3.3 Visual feature extraction module: the module applies 1×1 convolutions, group normalization, and SiLU activation to extract patch embeddings from visual data. These embeddings are passed through a series of H-MHSA, a MLP, and a downsampling operation to generate final visual features as output. 18

3.4 Transformer encoder architecture: Q, K, and V refer to input queries, keys, and values, respectively. 19

3.5 A brief overview of the MMFformer architecture for multimodal depression detection. The proposed approach comprises multiple modules: video feature extraction (top), audio feature extraction (middle), and late and intermediate fusion (bottom). Video feature extraction utilizes transformer blocks and residual connections to capture spatial patterns from video clips. Audio feature extraction involves processing audio signals with a transformer encoder to capture meaningful temporal dynamics in speech. The fusion module works at the late and intermediate stages to capture significant intermodal interactions among the extracted features. Finally, the fused features are fed into a classifier to detect depressive states from the multimodal inputs. 22

4.1 Visualization of features based on weights of the final embedding obtained by fusing audio-visual features; the darker the color, the higher the weights. 36

4.2 Feature visualization of the proposed MDD-Net through t-SNE for multimodal depression detection. 37

List of Tables

4.1 Dataset Feature Specifications for D-Vlog and LMVD. 31

4.2 Training Hyperparameters for MDD-Net and MMFformer. 33

4.3 Performances of Proposed MDD-Net for Depression Detection from Multimodal Vlog Data. 34

4.4 Performance Comparison of the Proposed MDD-Net with Different Existing Methods on D-Vlog Dataset. 35

4.5 Results of Ablation Studies for the Proposed MDD-Net Architecture to Detect Depression. 36

4.6 Performances of MMFformer for Depression Detection from Multimodal Vlog Data. Each Model is Run Ten Times to Obtain the Results (Mean ± Std). Bold Represents the Best and Underline Indicates the Second-best. 38

4.7 Comparison with the SOTA Methods on D-Vlog and LMVD Datasets for Depression Detection. 40

4.8 Results of Ablation Studies for MMFformer on the Video and Audio Data to Detect Depression. 41

4.9 Cross-corpus Validation Results between D-Vlog and LMVD Datasets on Multimodal Features. 41

A.1 Architecture Configuration of MDD-Net. 55

A.2 Architecture Configuration of MMFformer. 56

A.3 Comparative Summary of MDD-Net and MMFformer Configurations. 57

List of Abbreviations

A Audio 38

A+V Audio+Video 38

Acc Accuracy 33

AD-LSTM Attention-based Deep Long Short-term Memory 10

AD-MIL Attention-based Deep Multiple Instance Learning 10

Add Addition 35, 40

AFEM Acoustic Feature Extraction Module 4, 14, 16, 19

AI Artificial Intelligence 11, 28, 42

AST Audio Spectrogram Transformer 4, 6, 25, 32

BCE Binary Cross-Entropy 21

BDI Beck Depression Inventory 1

Bi-LSTM Bi-directional Long-Short Term Memory 8, 10, 34

CAIINET Contextual Attention and Information Interaction Mechanism 9, 10, 34

CNN Convolutional Neural Network 8, 10

Concat Concatenation 35, 40

CONCH CONtrastive learning from Captions for Histopathology 45

Conv1D 1D Convolution 26, 27

D-Vlog Depression-Vlog 3–5, 8–11, 13, 14, 29, 30, 33, 37–44

DepMSTAT Multimodal Spatio-Temporal Attention Transformer approach for Depression detection 11

EEG Electroencephalogram 8

eGeMAPS extended Geneva Minimalistic Acoustic Parameter Set 28–30, 44

F1 F1-Score 33

H-MHSA Hierarchical Multi-Head Self-Attention 3, 6, 18

IA Intermediate Attention 4, 13, 38, 41, 42

IoT Internet of Things 11

IT Intermediate Transformer 4, 13, 38, 39, 41, 42

LLM Large Language Model 12, 45

LMTformer Lightweight Multi-scale Transformer 11

LMVD Large-scale Multimodal Vlog Dataset 3–5, 10, 11, 13, 30, 38–42, 44

LSTM Long Short-Term Memory 8

LT Late Transformer 4, 13, 38, 39, 41, 42

MC Mutual Correlation 19

MDAVIF Multi-Domain Acoustical-Visual Information Fusion 9, 34

MDD-Net Multimodal Depression Detection Network 1, 3, 4, 6, 9, 13–15, 19, 28, 29, 32–35, 37, 42–45, 55, 57

MFCC Mel-Frequency Cepstral Coefficient 16, 29

MHSA Multi-Head Self-Attention 18

MIL Multiple Instance Learning 10

MLLM Multimodal Large Language Model 5, 12, 28, 42, 45

MLP Multi-Layer Perceptron 18, 24

MMFformer Multimodal Fusion Transformer Network 1, 4–6, 9, 13, 14, 21, 26, 28–30, 32, 33, 38–44, 56, 57

MT Mutual Transformer 4, 6, 13, 14, 19, 28, 35

MulT Multimodal Transformer 9

Multi Multiplication 35, 40

PHQ-9 Patient Health Questionnaire 1

Pr Precision 33

Rc Recall 33

RNN Recurrent Neural Network 3

SMT Spike Memory Transformer 11

SOTA State-of-the-Art 5–7, 12, 29, 39, 44

SSM State Space Model 10

STAT Spatial-Temporal Attentional Transformer 11

STST Spatio-Temporal Squeeze Transformer 11, 34

t-SNE t-Distributed Stochastic Neighbor Embedding 36, 37, 42

TAMFN Time-Aware Attention Multimodal Fusion Network 34

TCN Temporal Convolutional Network 8

TF Tensor Fusion 40

TFN Tensor Fusion Network 9, 34

UA Unweighted Average 33

UAA Unweighted Average Accuracy 38

UAF1 Unweighted Average F1-Score 38, 40, 41

UAP Unweighted Average Precision 38

UAR Unweighted Average Recall 38

V Video 38

VFEM Visual Feature Extraction Module 4, 14, 17

ViT Vision Transformer 4, 6, 23, 28, 32

VLM Vision Language Model 5, 12, 28, 42, 45

WA Weighted Average 33

WAA Weighted Average Accuracy 38, 40, 41

WAF1 Weighted Average F1-Score 38, 40, 41

WAP Weighted Average Precision 38, 40, 41

WAR Weighted Average Recall 38, 40, 41

WHO World Health Organization 1, 3

XGBoost eXtreme Gradient Boosting 10

Chapter 1

Introduction

This chapter introduces the problem of automated depression detection from audio-visual social media content, focusing on the role of transformer-based multimodal fusion in computational mental health analysis. It discusses the key challenges of extracting and combining acoustic and visual features from noisy, real-world video data and explains why existing approaches have struggled with these tasks. The scope of this research is defined through the development of two proposed architectures, Multimodal Depression Detection Network (MDD-Net) and Multimodal Fusion Transformer Network (MMFformer), each offering a distinct strategy for cross-modal feature fusion. Finally, the chapter outlines the unique contributions of this thesis in advancing automated depression screening from video blog data.

1.1 Problem Definition

Depression is one of the most common mental health conditions worldwide, affecting over 280 million people according to the World Health Organization (WHO) [1]. It disrupts daily functioning, diminishes quality of life, and in severe cases leads to self-harm or suicide. Despite its prevalence, diagnosing depression remains a difficult and largely subjective process. Clinicians typically rely on structured interviews and self-report questionnaires such as the Patient Health Questionnaire (PHQ-9) [2] or the Beck Depression Inventory (BDI) [3], which depend heavily on a patient's willingness and ability to accurately describe their emotional state. These tools are inherently limited: they cannot be easily administered at scale, they are influenced by social desirability bias, and they require access to trained professionals that many individuals simply do not have.

At the same time, millions of people voluntarily share personal video blogs on platforms like YouTube, where they talk about their daily lives, struggles, and emotional experiences. These videos contain rich behavioral signals, including facial expressions, voice quality, speech rhythm, and body language, that are difficult to consciously control and that often reveal more about a person's mental state than what they choose to say in words [4]. This has sparked growing research interest in leveraging social media data as a potential source of observable markers for depression [5, 6].

However, turning this idea into a working system is far from straightforward. Social media videos are recorded under uncontrolled conditions with varying lighting, background noise, camera angles, and video quality. Speakers may move in and out of frame, their faces may be partially occluded, and the audio may contain music, environmental sounds, or overlapping speech. These factors introduce significant noise, making feature extraction unreliable when standard pipelines are applied without careful consideration of the data characteristics [7].

A second, and perhaps more fundamental, problem lies in how to combine information across different modalities. Audio and video carry complementary information about depression: vocal features such as reduced pitch variation, slower speech rate, and lower energy often accompany depressive episodes [8], while visual cues like diminished facial expressiveness, reduced eye contact, and slower head movements provide additional evidence [9]. But these signals exist in different feature spaces, evolve at different temporal resolutions, and do not always align in time. Simply concatenating audio and video features, as many earlier methods have done, fails to preserve the directional relationships between modalities and often generates representations that are no better than using either modality alone [10].

Existing fusion strategies in the literature have largely relied on element-wise operations, concatenation, or tensor-based products [11] that treat the two modalities symmetrically and fail to model how one modality might inform or modulate the other. Attention-based methods have shown some improvement [12], but most of them apply attention within each modality separately without allowing explicit cross-modal interaction. This means the model cannot learn, for instance, that a particular facial expression becomes more predictive of depression when it co-occurs with a specific vocal pattern, or that changes in voice pitch carry different meanings depending on the accompanying visual context.

This thesis directly addresses these gaps. The main question is how to design transformer-based fusion architectures that can extract meaningful audio and visual features from noisy social media videos and combine them in a way that preserves and exploits the directional, complementary relationships between the two modalities for depression classification.

1.2 Motivation

The motivation behind this work stems from both clinical need and technical opportunity. On the clinical side, the gap between the number of people living with depression and those who actually receive a diagnosis remains enormous. The WHO estimates that in low and middle-income countries, more than 75% of people with mental health conditions receive no treatment at all [1]. Even in developed nations, long wait times, stigma, and lack of access to mental health professionals prevent many from seeking help [13]. By analyzing content individuals already share publicly, an automated screening tool could identify those at risk. This capability could serve as an early warning system and help direct limited clinical resources where they are needed most.

From a technical viewpoint, the transformer architecture [14] has fundamentally changed how sequential and multimodal data are processed. Unlike Recurrent Neural Networks (RNNs), which process sequences one step at a time and tends to lose information over long distances, transformers use self-attention to model relationships between all positions in a sequence simultaneously. This makes them particularly well-suited for tasks where long-range dependencies matter, such as capturing how a speaker's vocal tone shifts over several minutes in a video or how recurring facial micro-expressions build a pattern indicative of depression over time.

More importantly, the cross-attention mechanism within the transformer framework offers a natural way to model interactions between different modalities [12, 15]. By allowing audio and video features to interact with each other, the model can learn to identify which combinations of audio-visual patterns are most discriminative for depression classification. This goes beyond what earlier fusion methods could achieve with static operations like concatenation or element-wise multiplication [11].

The availability of curated benchmark datasets like Depression-Vlog (D-Vlog) [16] and Large-scale Multimodal Vlog Dataset (LMVD) [17], along with established low-level feature extraction tools such as OpenSMILE [18] for audio and Dlib [19] for facial landmarks, provides the infrastructure needed to develop and evaluate new fusion architectures. Together, these clinical and technical motivations drive the research presented in this thesis.

1.3 Scope of the Work

This research introduces two novel transformer-based multimodal fusion architectures for automated depression detection from social media video blogs. The first framework, MDD-Net, pairs a global self-attention network [20] for acoustic feature processing with a Hierarchical Multi-Head

Self-Attention (H-MHSA) [21] mechanism for visual feature processing and fuses the two streams through a Mutual Transformer (MT) that computes bidirectional cross-modal correlations. The second framework, MMFformer, takes a different architectural approach by employing a residual Vision Transformer (ViT) [22] for video and an Audio Spectrogram Transformer (AST) [23] for audio, and systematically compares three fusion strategies: Late Transformer (LT) fusion, Intermediate Transformer (IT) fusion, and Intermediate Attention (IA) fusion.

Both architectures are evaluated on the D-Vlog dataset [16], which serves as the primary benchmark throughout this thesis. The MMFformer is additionally tested on the LMVD dataset [17], and cross-corpus experiments are conducted to assess generalizability. The study is limited to binary depression classification using audio and visual modalities; text-based analysis of speech content is outside the scope of this work, as is clinical validation or deployment in real healthcare environments. The goal is to advance the technical foundations of multimodal fusion for this task, not to replace clinical judgment.

1.4 Contributions

The main contributions of this thesis are two-fold. In the first part, a robust MDD-Net is proposed in this work, which MT networks are deployed to efficiently capture and fuse acoustic and visual features. The proposed approach utilizes mutual transformers to compute correlations across modalities, providing a better representation of depression-related behavioral patterns using the D-Vlog database collected from social media platforms [16]. The following are the main contributions:

- A deep learning framework with transformer architecture is developed for depression detection through multimodal data.
- An Acoustic Feature Extraction Module (AFEM) is proposed, integrating a global self-attention mechanism to model content-based and positional relationships within the feature space.
- A Visual Feature Extraction Module (VFEM) is introduced to extract both local and global patterns through patch embeddings and hierarchical attention mechanisms.
- A MT is designed to calculate interdependencies between audio and visual embeddings, capturing multimodal information across modalities through joint representations.

- The D-Vlog dataset [16] is used to achieve cutting-edge performance, demonstrating the proposed model's effectiveness in classifying between depressed and non-depressed people based on non-verbal behavior.

Secondly, we present a multimodal fusion network, called MMFformer, for detecting depression from social media data. To address the challenges of extracting and fusing spatio-temporal information, we propose a system that captures high-level spatial features from video data using a transformer with residual connections, while simultaneously modeling the temporal dynamics of speech signals with a transformer encoder. Moreover, we propose a fusion module that incorporates late and intermediate fusion methods to enhance the relationship between modalities. For empirical experiments, we conduct comprehensive tests on depression datasets, D-Vlog [16] and LMVD [17], where MMFformer achieves superior results compared to State-of-the-Art (SOTA) approaches. Our major contributions are summarized as follows:

- A visual feature-extraction mechanism based on a residual learning transformer architecture is proposed, enabling the extraction of complex spatial patterns from dynamic facial expressions.
- An audio processing network employing a transformer encoder is developed to effectively preserve temporal dependencies in speech relevant to depression signals.
- A fusion module comprising late transformer fusion, intermediate transformer fusion, and intermediate attention fusion is introduced to improve the interaction between audio and visual modalities.
- Comprehensive tests using two publicly accessible datasets, D-Vlog and LMVD, illustrate that our developed system outperforms several current SOTA methods.

1.5 Thesis Organization

The rest of this thesis is organized as follows:

Chapter 2 reviews the existing literature on depression detection, covering traditional clinical approaches, unimodal methods based on audio or video alone, and multimodal fusion strategies that have been proposed in recent years. The chapter also examines the evolution of transformer architectures and their application to affective computing tasks, including recent developments in Vision Language Models (VLMs) and Multimodal Large Language Models (MLLMs), and identifies the specific gaps this thesis aims to address.

Chapter 3 presents the detailed architectures of the two proposed frameworks. The first part describes MDD-Net, including the global self-attention network for acoustic features, the H-MHSA for visual features, the MT fusion module, and the composite loss function. The second part introduces MMFformer, covering the residual ViT, the AST, and the three fusion strategies.

Chapter 4 describes the datasets, experimental setup, and evaluation metrics used to assess both frameworks. It reports the benchmarking performance of MDD-Net and MMFformer relative to SOTA methods, presents ablation studies examining the contribution of each architectural component, and discusses cross-corpus validation results. The chapter also includes a comparative analysis of the different fusion strategies explored in this thesis.

Chapter 5 concludes the thesis by summarizing the key findings, discussing the limitations of the current work, and suggesting directions for future research.

Chapter 2

Literature Review

This chapter reviews the existing literature on depression detection, covering approaches that range from traditional clinical methods to modern deep learning and transformer-based architectures. The review is organized into three main areas. The first covers deep learning approaches for depression detection, including unimodal methods that operate on a single data source, multimodal methods that combine multiple modalities, and audio-visual fusion strategies that have been specifically applied to combine acoustic and visual information. The second examines advanced multimodal architectures proposed for depression detection in recent years, with particular emphasis on SOTA and attention-based designs. The third reviews transformer-based approaches and their growing role in affective computing tasks. The chapter concludes by identifying the specific gaps in the current literature that this thesis aims to address.

2.1 Depression Detection Using Deep Learning

Recent years have shown substantial advances in multimodal depression detection through deep learning approaches applied to vlog data. Current research emphasizes the use of deep learning for depression detection, which is more effective compared to manual feature extraction [24]. Some researchers have used single-modality data for depression recognition, while others have used multimodal data that contain comprehensive information for accurate and reliable depression detection [25]. The following subsections outline relevant research on methodologies applied to depression detection, including both unimodal and multimodal approaches using various data modalities.

- **Unimodal Depression Detection**

In the domain of unimodal depression detection, significant research has been conducted across various modalities to assess depressive symptoms. In [26], natural language processing is

used to detect depressive symptoms by analyzing linguistic patterns and emotional expressions in social media text extracted from platforms such as X (formerly Twitter). In audio-based approaches, several techniques have been explored. Low-level descriptors and spectrogram images are commonly used as inputs to various neural network architectures to predict depression levels. For example, Kim et al. [27] have exploited log-Mel spectrograms with a Convolutional Neural Network (CNN) architecture to predict depression. Electroencephalogram (EEG) brain data, often used in brain-computer interface applications, has also been used by some researchers for emotion recognition tasks [28]. In the video-based domain, researchers have applied various visual features and deep learning techniques to detect depression. For instance, Zhou et al. [29] have introduced DepressNet, a deep regression network designed to learn visually clear representations for depression recognition from facial expressions. The system generates a depression activation map that highlights facial regions indicative of depression severity, improving both accuracy and interpretability in depression detection.

- **Multimodal Depression Detection**

Recent advancements focus on multimodal feature fusion, combining text, audio, and visual data to enhance depression detection [30]. Video content from social media platforms such as YouTube has been used to detect depression by analyzing both audio and visual features. A technique [31] fusing these features from video logs achieved over 75% accuracy in recognizing depressive symptoms. In another work [32], researchers introduced a method for depression detection that combines acoustic and visual data via a time-aware attention mechanism and a Temporal Convolutional Network (TCN), using the D-Vlog dataset. The system proposed a framework integrating Long Short-Term Memory (LSTM) with contextual attention and global information interaction. The proposed architecture used local information fusion to combine temporal acoustic and visual features, achieving a precision of 66.56% on the D-Vlog dataset.

Seneviratne et al. [33] presented a multimodal depression detection architecture combining audio and text data, addressing overfitting by employing multiple levels of classifiers. Yuan et al. [34] introduced a multi-order factor fusion technique to take advantage of high-order interactions between different kinds of modalities. Shen et al. [35] developed a depression detection system using Bi-directional Long-Short Term Memory (Bi-LSTM) and fused audio and text data with attention mechanisms.

- **Audio-Visual Fusion Strategies**

Since this thesis specifically targets audio-visual feature fusion for depression screening, it is important to review how prior works have combined these two modalities. Audio-visual fusion

strategies in the literature can broadly be categorized into early fusion, late fusion, and hybrid approaches. Early fusion methods concatenate raw or low-level features from both modalities before feeding them into a shared classifier. While straightforward, early fusion often suffers from the dimensionality mismatch between audio and visual feature spaces because acoustic descriptors and facial landmarks operate at different temporal resolutions and carry fundamentally different types of information.

In contrast, late fusion methods process each modality independently using separate networks and combine decision-level outputs or high-level representations at the final stage. The Depression Detector proposed alongside the D-Vlog dataset [16] adopted a cross-transformer that performed late-stage fusion of audio and visual embeddings. Multi-Domain Acoustical-Visual Information Fusion (MDAVIF) [36] employed two autoencoders to preserve modality-specific information before combining them, achieving an F1-Score of 75.25% on D-Vlog. These methods retain modality-specific patterns but limit the opportunity for audio and visual features to interact during representation learning.

Intermediate or feature-level fusion attempts to bridge this gap by allowing cross-modal interaction at hidden layers. The Tensor Fusion Network (TFN) [11] computes an outer product of unimodal representations to model higher-order interactions, but its quadratic growth in parameters makes it impractical for longer sequences. Attention-based fusion methods, such as Contextual Attention and Information Interaction Mechanism (CAIINET) [37] discussed further in Section 2.2, apply attention weights within each modality but do not enable one modality to directly attend to the other. The Multimodal Transformer (MulT) [12] introduced directional cross-modal attention for unaligned multimodal sequences, but it was designed for sentiment analysis and has not been directly applied to vlog-based depression detection.

In the literature, several studies have attempted to fuse audio and visual modalities for depression detection; however, existing frameworks still face notable difficulties in identifying the most discriminative patterns from raw multimodal data and in developing fusion approaches that go beyond symmetric operations. None of the current audio-visual fusion methods simultaneously provides bidirectional cross-modal attention, where audio queries attend to visual features and visual queries attend to audio features, combined with a joint transformer that further refines the fused representation. This particular gap limits the ability of existing architectures to capture directional and complementary relationships between acoustic and visual cues, which are essential for distinguishing depressive behavior in uncontrolled social media environments. The mutual transformer in MDD-Net and the systematic three-level fusion comparison in MMFformer are designed to address this shortcoming, as detailed in Chapter 3.

2.2 Advanced Multimodal Architectures for Depression Detection

DepMamba [38] introduced an audio-visual progressive fusion network based on Mamba to detect depression through multiple data modalities. The architecture combined CNNs and Mamba to capture local-to-global features across long-range sequences. It incorporated a multimodal collaborative State Space Model (SSM) to extract both intermodal and intramodal information for each modality. A multimodal-enhanced SSM is employed to further strengthen the cohesion between modalities. Experimental results showed that DepMamba achieved an accuracy of 68.87% on the D-Vlog dataset and 72.13% on the LMVD dataset. Xing et al. [39] presented a multimodal depression detection framework, EMO-Mamba, that leveraged multimedia data to improve performance. The technique applied a CNN to extract the spatial attributes of facial features and the local acoustic features from audio. The SSM network is utilized to understand temporal variations and efficiently maintain memory over long time-series. A multimodal fusion framework is proposed to efficiently integrate crucial information from multiple modalities, thereby enhancing overall detection capabilities. In the D-Vlog dataset, EMO-Mamba obtained accuracy, precision, recall, and F1-Score of 75.54%, 75.79%, 75.54%, and 75.66%, respectively. Shangguan et al. [40] proposed a Multiple Instance Learning (MIL) approach for detecting depression using social media data. The proposed MIL architecture is designed to handle long-term sequences of visual data through an Attention-based Deep Long Short-term Memory (AD-LSTM) network. The AD-LSTM processed fixed-length visual and speech segments to retrieve temporal dimensions of each instance, and the Attention-based Deep Multiple Instance Learning (AD-MIL) block fused the temporal representations obtained to perform depression detection. Compared with existing benchmarks, the experiments demonstrated that the developed MIL network achieved the highest weighted average precision, recall, and F1-Score of 67.27%, 67.77%, and 66.64%, respectively. In another research, Zhou et al. [37] developed a deep learning based framework called CAIINET for the early detection of depression, utilizing contextual attention and an information interaction mechanism. The proposed system used a contextual attention module with a Bi-LSTM model to capture crucial audio and visual cues at important temporal points. The system incorporated local and global information fusion modules that evaluated the significance and interactions among the extracted attributes at both levels. Experiments on the D-Vlog dataset revealed that CAIINET surpassed current benchmark models, achieving 66.56%, 66.98%, and 66.55% for weighted average precision, recall, and F1-Score, respectively. Kowalewski et al. [41] compared machine learning and deep learning techniques for depression detection using audiovisual social media content. The proposed approach applied three learning algorithms, including EfficientNet, neural network, and eXtreme Gradient Boosting (XGBoost) on D-Vlog dataset and obtained the highest F1-Score of 77% from the XGBoost classifier.

2.3 Transformer-Based Approaches for Depression Detection

He et al. [42] developed a lightweight architecture named Lightweight Multi-scale Transformer (LMTformer), aimed to detect depression from facial videos through a multi-scale transformer. This model retrieved coarse-grained attributes from facial expressions and then processed them through a lightweight multi-scale transformer. The transformer recorded local and global patterns in diverse receptive fields. Moreover, global features are enhanced by a multi-scale global feature fusion approach. Using the LMVD dataset, the proposed network achieved accuracy, precision, recall and F1-Score of 82.76%, 82.87%, 82.76% and 82.74%, respectively. Further, a video-based depression detection system termed Depressformer is introduced in [43]. The model used the Swin Transformer to enhance the extraction of key video features. A module focused on depression-specific fine-grained local feature extraction is presented to identify detailed signs of depression. In addition, a depression channel attention fusion block is added to improve the fusion and modeling of the combined features. The empirical findings demonstrated its performance, obtaining an F1-Score of 0.59 on the D-Vlog dataset. Tao et al. [44] proposed a depression detection model called Multimodal Spatio-Temporal Attention Transformer approach for Depression detection (DepMSTAT), which analyzes audio and visual features from vlog content to identify depression. The Spatial-Temporal Attentional Transformer (STAT) block is at the core of the system, designed to effectively capture spatial and temporal relationships within multimodal data. This module extracted spatio-temporal features from individual modalities and fused them to analyze vlog-based audio and visual signals. According to experimental results, DepMSTAT achieved precision of 71.53%, recall of 75.60%, and F1-Score of 73.51%. Further, Tao et al. [45] presented a Spatio-Temporal Squeeze Transformer (STST) technique to extract relevant semantic features related to depression. The approach employed a transformer encoder to process spatiotemporal data and extract significant features, which were then used by a voting-based classifier to detect depression. The experiments on the D-Vlog dataset achieved an accuracy of 70.70%, a precision of 72.50%, a recall of 77.67%, and an F1-Score of 75%. Yang et al. [46] deployed a computationally efficient hierarchical structure for autonomous depression detection in an Internet of Things (IoT) environment. This framework enabled IoT devices to collaborate in a layered and distributed way to obtain mental health information. The proposed method trained the Spike Memory Transformer (SMT) to capture complex temporal relationships and heterogeneous patterns within data for depression recognition. Experimental results showed that SMT outperformed traditional deep learning methods with an accuracy of 70.73% for D-Vlog dataset and obtained lower consumption of energy during inference. While the above methods represent significant progress in applying transformers to depression detection, the field of multimodal Artificial Intelligence (AI) has been evolving even more rapidly in parallel.

Beyond task-specific transformer architectures, the broader landscape of multimodal AI has

been reshaped by the emergence of VLMs and MLLMs that process and reason across text, image, audio, and video within unified frameworks. In affective computing, Cheng et al. [47] proposed Emotion-LLaMA, a MLLM that integrates audio, visual, and textual inputs through emotion-specific encoders, achieving SOTA results on multiple emotion recognition benchmarks. Lian et al. [48] introduced AffectGPT with pre-fusion operations for enhanced multimodal emotion understanding. For mental health specifically, Yang et al. [49] developed MentaLLaMA, the first open-source instruction-following Large Language Model (LLM) for interpretable mental health analysis on social media, covering eight tasks including depression detection. Chen et al. [50] transformed clinical interviews into graph structures and used GPT-based data augmentation for depression detection, while Zhang et al. [51] integrated acoustic speech landmarks into Llama2 for audio-based depression screening. More recently, Sadeghi et al. [52] demonstrated a fully automated pipeline combining LLM-extracted depression indicators from interview transcripts with facial expression features for multimodal depression detection. Although these VLM and MLLM approaches have not yet been applied to vlog-based audiovisual depression screening of the kind addressed in this thesis, they represent a promising direction that could complement the transformer-based fusion strategies proposed in the current work.

2.4 Summary and Research Gaps

The literature reviewed in this chapter reveals several important trends and limitations in the current state of automated depression detection. First, while unimodal methods have demonstrated that individual modalities such as audio and video carry meaningful signals for depression, they consistently underperform compared to multimodal approaches. This indicates that depression manifests through a combination of behavioral cues that cannot be fully captured by any single modality alone.

Secondly, as discussed in *Audio-Visual Fusion Strategies*, existing multimodal fusion strategies predominantly rely on symmetric operations that fail to model directional relationships between audio and visual features, and none of the current approaches provide bidirectional cross-modal attention combined with joint transformer refinement.

Third, although transformers have shown remarkable success in various sequence modeling tasks, their application to multimodal depression detection remains limited. Most transformer-based methods in this domain apply self-attention within each modality separately, without enabling explicit cross-modal interaction. This means the models cannot learn how specific combinations of audio and visual patterns jointly indicate depression.

Finally, comparing existing methods remains difficult because different studies report results using different evaluation metrics and follow different experimental protocols, making direct

benchmarking unreliable. The D-Vlog dataset [16] has emerged as a standard benchmark, but prior methods have reported F1-Scores between 59% and 77%, suggesting substantial room for improvement. The LMVD dataset [17] provides a complementary evaluation platform but has received limited attention in the literature. Moreover, the unique challenges of social media content, including uncontrolled recording conditions, varying audio quality, and inconsistent lighting, require architectures robust enough to extract reliable features from noisy real-world data.

These gaps motivate the two architectures proposed in this thesis. MDD-Net addresses the fusion limitation through a MT that computes bidirectional cross-modal correlations between acoustic and visual embeddings, while MMFformer systematically compares three different fusion strategies, including LT, IT, and IA fusion to determine which level of cross-modal interaction is most effective for depression detection from social media video blogs.

Chapter 3

Proposed Methodology

This chapter presents the detailed architectures of the two proposed frameworks for multimodal depression detection. The first part describes MDD-Net, a multimodal depression detection network that employs a MT for bidirectional cross-modal fusion. The second part introduces MMFformer, a multimodal fusion transformer network that systematically compares three fusion strategies for combining audio and visual features. Both architectures are designed to extract meaningful representations from social media video data and fuse them to detect depression.

3.1 MDD-Net: Multimodal Depression Detection through Mutual Transformer

The first proposed framework, MDD-Net, is a Multimodal Depression Detection Network where MT networks are deployed to efficiently capture and fuse acoustic and visual features. The proposed approach utilizes MTs to compute correlations across modalities, providing a better representation of depression-related behavioral patterns using the D-Vlog database collected from social media platforms [16].

MDD-Net consists of four core modules: an AFEM for retrieving relevant acoustic attributes, a VFEM for extracting significant high-level patterns, a MT for computing the correlations among the generated features and fusing these features from multiple modalities, and a detection layer for detecting depression using the fused feature representations.

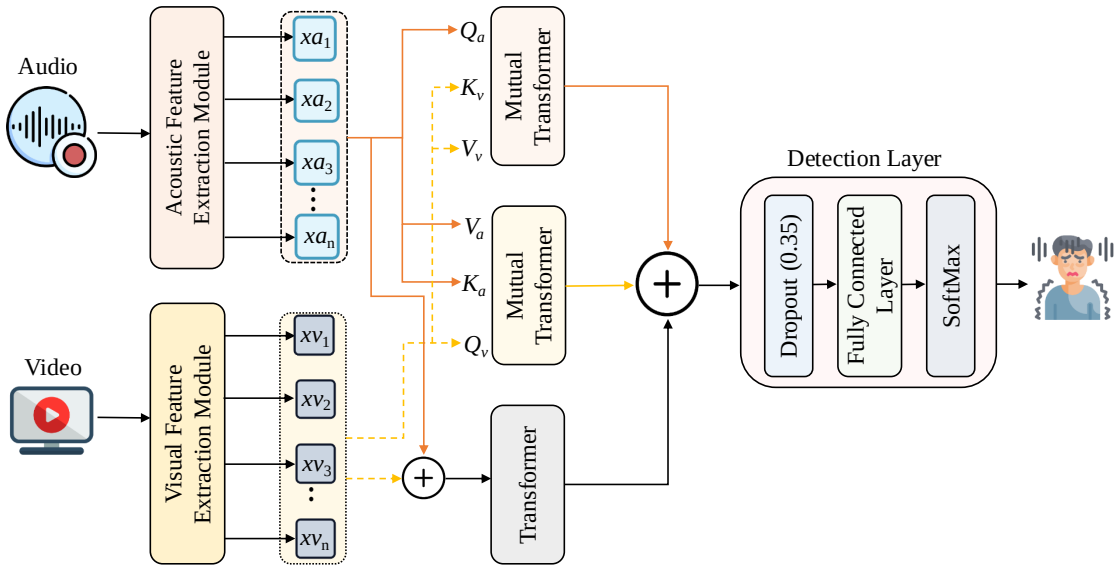


Figure 3.1: The overall architecture of the proposed MDD-Net system: a multimodal fusion technique for audio-visual feature fusion using mutual transformers for depression detection.

3.1.1 Problem Formulation

The depression detection problem in this work is formulated as a binary classification task. Let a multimodal vlog dataset be defined as $G = \{(X_A^n, X_V^n)\}_{n=1}^{|G|}$, where X_A^n are acoustic and X_V^n are visual features. The objective is to produce classifications $C = \{c_n\}_{n=1}^{|G|}$, where $c_n \in \{\text{Depression}, \text{Normal}\}$.

The dataset is formalized as follow:

$$G = \left\{\left(X_A^n, X_V^n\right)\right\}_{n=1}^{|G|}, \quad X_A^n \in \mathbb{R}^{t \times d_a}, \quad X_V^n \in \mathbb{R}^{t \times d_v} \tag{3.1}$$

Here, X_A^n refers to the acoustic features and X_V^n represents the visual features. The dimensions of the acoustic and visual features are represented by d_a and d_v. The sequence length of the data is t. Both the acoustic and visual sequences are aligned to have the same length. The objective is to classify each vlog in G as "Depression" or "Normal" by learning important patterns from the vlog's acoustic and visual features. The overall architecture of the proposed MDD-Net is illustrated in Fig. 3.1. The step-by-step learning process of the proposed MDD-Net is systematically described in Algorithm 1, which outlines how the acoustic and visual features are extracted, fused through the mutual transformer, and used for depression classification.

Algorithm 1 Multimodal Depression Detection Network

Input: Vlog dataset $G = \{(X_A^n, X_V^n)\}_{n=1}^{|G|}$
where X_A^n are acoustic and X_V^n are visual features
Output: Classifications $C = \{c_n\}_{n=1}^{|G|}$
where $c_n \in \{\text{Depression}, \text{Normal}\}$
for each vlog (X_A^n, X_V^n) in G **do**
    Extract Acoustic Features:
    $X_A^o \leftarrow \text{AFEM}(X_A^n)$ ▷ *Process acoustic features using global self-attention for content and positional relationships.*
    Extract Visual Features:
    $X_V^o \leftarrow \text{VFEM}(X_V^n)$ ▷ *Process visual features with patch embedding and transformer blocks.*
    Fuse Features with Mutual Transformer:
    $MC_{AV} \leftarrow \text{MutualTransformer}_{AV}(X_A^o, X_V^o)$ ▷ *Compute audio-to-video correlation.*
    $MC_{VA} \leftarrow \text{MutualTransformer}_{VA}(X_V^o, X_A^o)$ ▷ *Compute video-to-audio correlation.*
    $MC_{f_{AV}} \leftarrow \text{TransformerFusion}(X_A^o, X_V^o)$ ▷ *Fuse acoustic and visual features jointly.*
    $Z \leftarrow \text{Combine}(MC_{AV}, MC_{VA}, MC_{f_{AV}})$ ▷ *Generate multimodal representation.*
    Detect Depression:
    $p \leftarrow \text{DetectionLayer}(Z)$ ▷ *Compute probability via attention mechanism.*
    if $p > 0.5$ **then**
        $c_n \leftarrow \text{Depression}$
    else
        $c_n \leftarrow \text{Normal}$
    end if
end for
return C

Each of the proposed system modules is detailed in the following sections.

3.1.2 Acoustic Feature Extraction Module

In the AFEM, a combination of a global self-attention network [20] is used as the backbone for extracting relevant features from acoustic data. This module takes 25 low-level acoustic descriptors extracted from each audio segment, including loudness, Mel-Frequency Cepstral Coefficients (MFCCs), and spectral flux as inputs. The major components of the audio feature extraction module are represented in Fig. 3.2. Let, $X_A^i \in \mathbb{R}^{t \times d_a}$ and $X_A^o \in \mathbb{R}^{t \times d_o}$ represent the input and output feature vectors, respectively. The output feature vector is generated by fusing input data using both content-based and positional attention layers. The matrices of keys, queries, and values are defined as $K_a = [k_{ij}] \in \mathbb{R}^{t \times d_a}$, $Q_a = [q_{ij}] \in \mathbb{R}^{t \times d_a}$, and $V_a = [v_{ij}] \in \mathbb{R}^{t \times d_o}$, which are generated by applying three 1×1 convolutions on the input feature map X_A^i. The generated feature map $X_A^c \in \mathbb{R}^{t \times d_a}$ is then computed using content-based dot-product attention as in (3.2).

$$X_A^c = Q_a \cdot \rho(K_a^\top) \cdot V_a \tag{3.2}$$

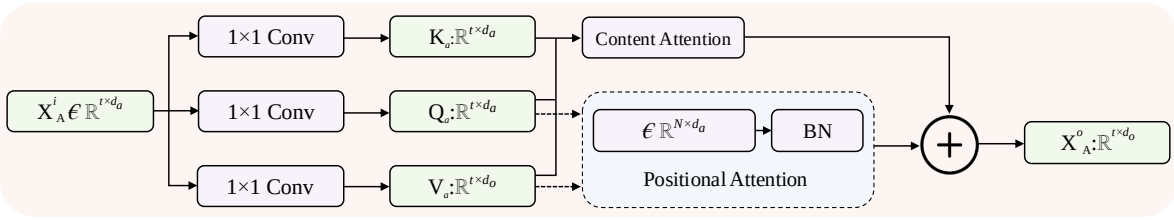


Figure 3.2: Acoustic feature extraction module: the module processes input audio features using 1×1 convolutions, applies both content and positional attention mechanisms, and performs batch normalization before fusing the features for the final representations.

where, ρ represents Softmax normalization. This attention layer first aggregates the elements of V_a into d_a global context vectors, weighted by $\rho(K_a^\top)$, and then redistributes the context vectors to individual elements based on the weights in Q_a. The Softmax normalization on queries is removed to allow the output features to span the entire subspace of the context vectors d_a.

To address the lack of spatial awareness in the content attention layer, a positional attention layer is introduced to handle both content and spatial relationships, as proposed in [53]. This layer analyzes the feature space in certain directions, ensuring that each element interacts with its neighbors. This approach effectively spreads information across the entire $N \times N$ neighborhood. Let, $\Delta = \left\{-\frac{N-1}{2}, \ldots, 0, \ldots, \frac{N-1}{2}\right\}$ be a set of N offsets, and $R = [r_\delta] \in \mathbb{R}^{N \times d_a}$ denotes the matrix of learnable relative position embeddings, where $\delta \in \Delta$. Then, our attention mechanism uses relative position embeddings (R) as keys, as shown in (3.3).

$$X^p_{A(xy)} = Q_{a(xy)} \cdot R^\top \cdot V_{a(xy)} \tag{3.3}$$

where, $Q_{a(xy)}$ is the query, $V_{a(xy)} = [v_{x+\delta,y}] \in \mathbb{R}^{N \times d_o}$ is the matrix including the values at the N neighboring elements of position (x, y), and $X^p_{A(xy)}$ refer to the output of the positional attention mechanism.

The final output feature map (X^o_A) is a combination of these two attention mechanisms, as represented in (3.4). Here, f_{bn} is the 1×1 convolution with Batch Normalization (BN) layer.

$$X^o_A = X^c_A + f_{bn}(X^p_A) \tag{3.4}$$

3.1.3 Visual Feature Extraction Module

For VFEM, the network uses a patch embedding layer followed by a transformer block. The inputs of this module are 68 extracted facial landmarks for each frame in the vlog, which are fed as 136-

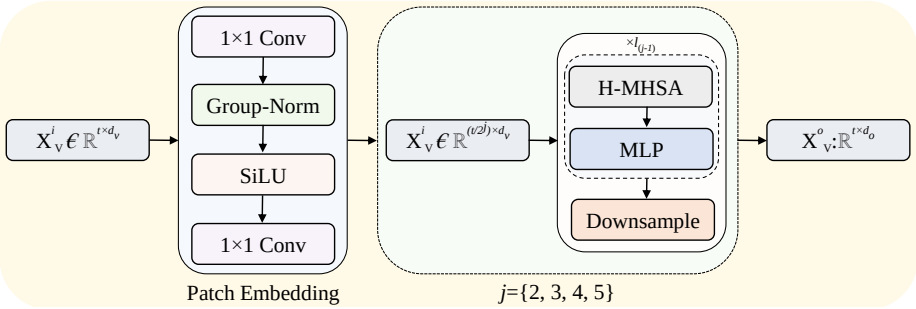


Figure 3.3: Visual feature extraction module: the module applies 1×1 convolutions, group normalization, and SiLU activation to extract patch embeddings from visual data. These embeddings are passed through a series of H-MHSA, a MLP, and a downsampling operation to generate final visual features as output.

dimensional vectors. The input feature map, $X_V^i \in \mathbb{R}^{t \times d_v}$ is processed through a 1×1 convolution layer, followed by group normalization, a SiLU activation function, and another 1×1 convolution. This acts as the patch embedding mechanism, projecting the input data into a suitable dimension for the transformer block. The video feature extraction module is represented in Fig. 3.3.

Once the embedding is complete, the data is passed through transformer blocks. In the transformer block, Query (Q_v), Key (K_v), and Value (V_v) are calculated using (3.5).

$$Q_v = X_V^i \beta^q, \quad K_v = X_V^i \beta^k, \quad V_v = X_V^i \beta^v \tag{3.5}$$

where, $\beta^q, \beta^k, \beta^v \in \mathbb{R}^{d_v \times d_v}$ belong to the weight matrices of linear transformations during training. The H-MHSA [21] is formulated in (3.6) considering the same dimensions for input and output.

$$M_V = \rho \left(\frac{Q_v K_v^\top}{\sqrt{d}} \right) V_v \tag{3.6}$$

After computing Multi-Head Self-Attention (MHSA), a residual connection is incorporated to enhance optimization, as mentioned in (3.7).

$$M_V' = M_V \beta_p + X_V^i \tag{3.7}$$

where, $\beta_p \in \mathbb{R}^{d_v \times d_v}$ is a trainable weight matrix for feature projection. A Multi-Layer Perceptron (MLP) is applied to improve the representation, as depicted in (3.8).

$$X_V^o = f_{\text{mlp}}(M_V') + M_V' \tag{3.8}$$

where, X_V^o represents the output of the transformer module. The transformer module is repeated $l_{(j-1)}$ times, where $j = \{2, 3, 4, 5\}$.

3.1.4 Mutual Transformer

The proposed approach uses the MT module to calculate multimodal mutual correlations and speaker-related depression attributes. The block diagram of the transformer is demonstrated in Fig. 3.4. From the AFEM, the acoustic features are obtained: $X_A^o = \{x_{a1}, x_{a2}, \ldots, x_{an}\}$, where $n = 256$ refers to the sequence length and $d_a = 71$ specifies the size for each embedding. Similarly, the visual feature extraction module provides the visual feature $X_V^o = \{x_{v1}, x_{v2}, \ldots, x_{vn}\}$, where $n = 256$ is the length of the sequence and $d_v = 139$ is the embedding size. The MT module bidirectionally calculates audio-to-video (MC_{AV}) and video-to-audio (MC_{VA}) correlations, followed by a joint transformer fusion.

The Mutual Correlation (MC) is obtained using two mutual transformer layers: $MC_{\text{AV}} \in \mathbb{R}^{n \times d}$ and $MC_{\text{VA}} \in \mathbb{R}^{n \times d}$ from $X_A^o \in \mathbb{R}^{n \times d_a}$ and $X_V^o \in \mathbb{R}^{n \times d_v}$. Another transformer layer is used to fuse both $X_A^o \in \mathbb{R}^{n \times d_a}$ and $X_V^o \in \mathbb{R}^{n \times d_v}$. The mutual transformers receive query, key, and value from different modalities. In our MDD-Net, the MC_{AV} mutual transformer receives the audio embedding X_A^o as Q_a and video embedding X_V^o as K_v and V_v. Similarly, MC_{VA} receives X_V^o as Q_v and X_A^o as K_a and V_a. At the same time, another transformer is applied to extract the relevant features $MC_{f_{\text{AV}}} \in \mathbb{R}^{(2n) \times (d_a + d_v)}$ through the fusion of X_A^o and X_V^o.

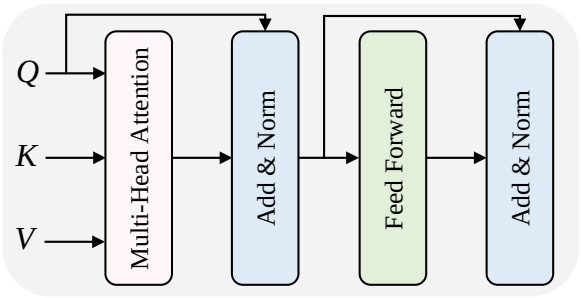


Figure 3.4: Transformer encoder architecture: Q, K, and V refer to input queries, keys, and values, respectively.

The corresponding formulations are presented in (3.9), (3.10), and (3.11).

$$
\begin{aligned}
f_{\text{attn}(X_A^o, X_V^o)} &= \rho\left(\frac{Q_a K_v^\top}{\sqrt{d_{K_v}}}\right) V_v, \\
f_{\text{attn}(X_V^o, X_A^o)} &= \rho\left(\frac{Q_v K_a^\top}{\sqrt{d_{K_a}}}\right) V_a
\end{aligned}
\tag{3.9}
$$

$$
\begin{aligned}
MC_{\text{AV}} &= L\left(X_A^o + \sigma\left(\beta^\top\left(L(X_A^o + f_{\text{attn}(X_A^o, X_V^o)})\right)\right) + b\right), \\
MC_{\text{VA}} &= L\left(X_V^o + \sigma\left(\beta^\top\left(L(X_V^o + f_{\text{attn}(X_V^o, X_A^o)})\right)\right) + b\right)
\end{aligned}
\tag{3.10}
$$

$$
MC_{f_{\text{AV}}} = f_{\text{transformer}}(f_{\text{concat}}(X_A^o, X_V^o)) \tag{3.11}
$$

where, L denotes the LayerNorm. f_{attn} is the self-attention function (as represented in (3.9)), $f_{\text{transformer}}$ and f_{concat} are the transformer and concatenation layer function of X_A^o and X_V^o. β and b are learnable parameters.

The multimodal transformer representation $Z \in \mathbb{R}^{(4n)\times 2(d_a+d_v)}$ is obtained using mutual transformer fusion technique. The fused output in the form of Z is shown in (3.12).

$$
Z = f_{\text{concat}}\left(f_{\text{AvgPool}}\left(MC_{\text{AV}}, MC_{\text{VA}}, MC_{f_{\text{AV}}}\right)\right) \tag{3.12}
$$

3.1.5 Depression Detection Layer

The output class probabilities for multimodal depression detection are derived as in (3.13).

$$
\begin{aligned}
\alpha &= F(\beta^\top \cdot Z), \\
p &= \rho\left(\beta^\top(\alpha^\top \cdot Z) + b\right)
\end{aligned}
\tag{3.13}
$$

Here, $F(\cdot)$, α, and $p \in \mathbb{R}^2$ denote the fully connected layer, attention scores, and depression detection probability, respectively.

3.1.6 Loss Function

For depression detection, a customized loss function is used that combines multiple losses to address the challenges of noisy and imbalanced datasets [54]. The customized loss function used in this paper combines Binary Cross-Entropy (BCE), Focal Loss, and L2 regularization, as shown in (3.14).

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Focal}} + \mathcal{L}_{\text{L2}} \tag{3.14}$$

The binary cross-entropy is expressed mathematically by (3.15), where y is the true class label and ϵ_s is the smoothing factor.

$$\begin{aligned} y_s &= y \cdot (1 - \epsilon_s) + 0.5 \cdot \epsilon_s, \\ \mathcal{L}_{\text{BCE}} &= -\left(y_s \log(p) + (1 - y_s) \log(1 - p)\right) \end{aligned} \tag{3.15}$$

Moreover, the Focal Loss is defined as in (3.16), where ϕ is a scaling factor and γ is the focusing parameter.

$$\mathcal{L}_{\text{Focal}} = \phi \cdot (1 - p)^{\gamma} \cdot \mathcal{L}_{\text{BCE}} \tag{3.16}$$

Furthermore, the L2 regularization loss is computed as shown in (3.17), where θ_i represents the network parameters and λ is the strength of the L2 regularization.

$$\mathcal{L}_{\text{L2}} = \lambda \sum \|\theta_i\|_2^2 \tag{3.17}$$

3.2 MMFformer: Multimodal Fusion Transformer Network for Depression Detection

The second proposed framework, MMFformer, is a multimodal depression-detection network designed to extract high-level spatio-temporal patterns from multimodal social media data. The transformer network with residual connections captures spatial features from videos, and a transformer encoder is exploited to design important temporal dynamics in audio. Moreover, the fusion architecture combined the extracted features using late and intermediate fusion strategies to identify the most relevant intermodal correlations. The overall architecture of MMFformer is illustrated in Fig. 3.5. The complete processing pipeline of MMFformer, including the three fusion strategies, is presented in Algorithm 2. The architecture's individual modules are described in the following sections.

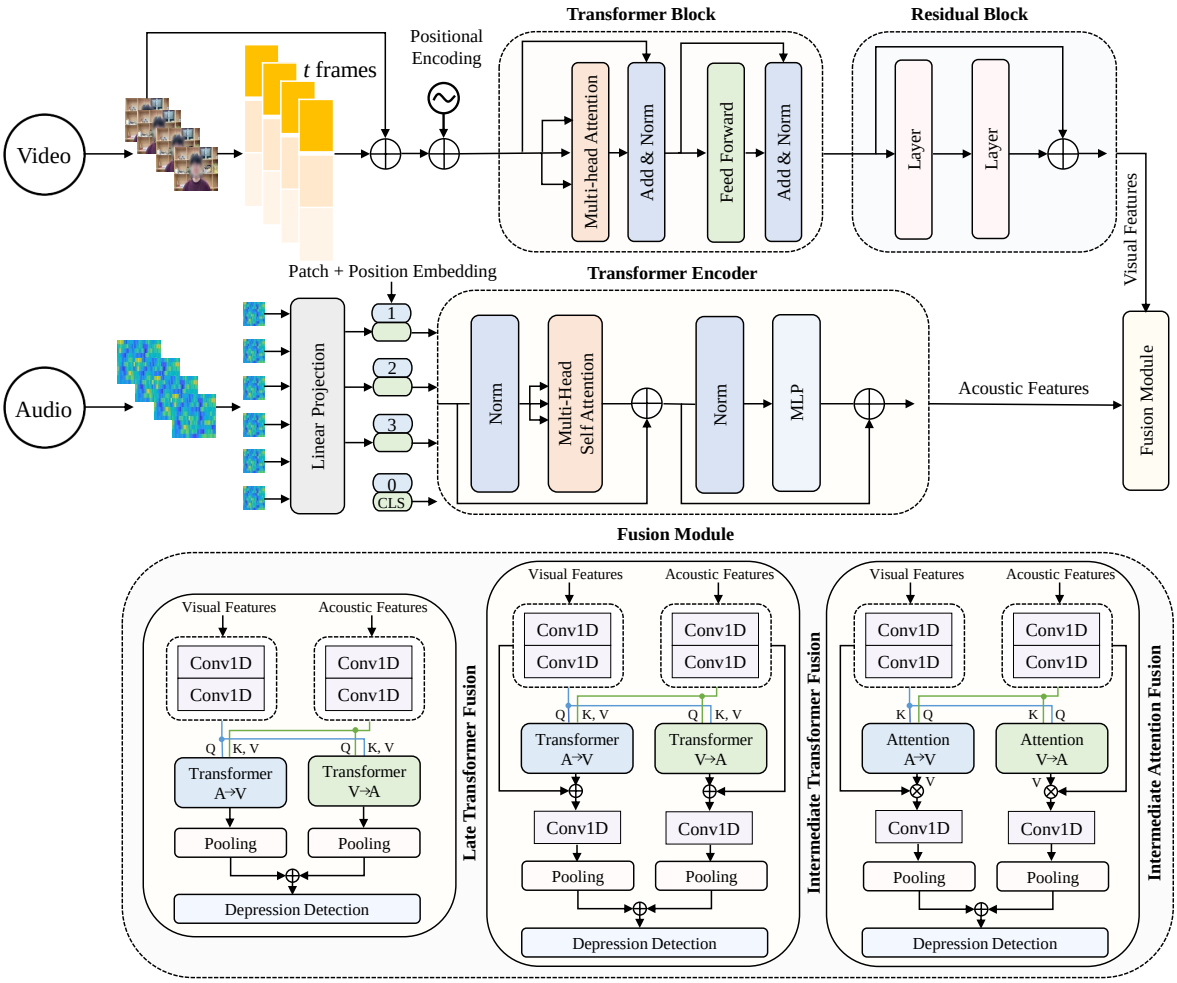


Figure 3.5: A brief overview of the MMFformer architecture for multimodal depression detection. The proposed approach comprises multiple modules: video feature extraction (top), audio feature extraction (middle), and late and intermediate fusion (bottom). Video feature extraction utilizes transformer blocks and residual connections to capture spatial patterns from video clips. Audio feature extraction involves processing audio signals with a transformer encoder to capture meaningful temporal dynamics in speech. The fusion module works at the late and intermediate stages to capture significant intermodal interactions among the extracted features. Finally, the fused features are fed into a classifier to detect depressive states from the multimodal inputs.

Algorithm 2 Multimodal Fusion Transformer Network for Depression Detection

Input: Vlog dataset $G = \{(\mathcal{X}_v^n, \mathcal{X}_a^n)\}_{n=1}^{|G|}$
where $\mathcal{X}_v^n$ are visual and $\mathcal{X}_a^n$ are audio features
Input: Fusion strategy $S \in \{\text{LT}, \text{IT}, \text{IA}\}$
Output: Classifications $C = \{c_n\}_{n=1}^{|G|}$
where $c_n \in \{\text{Depression}, \text{Normal}\}$
for each vlog $(\mathcal{X}_v^n, \mathcal{X}_a^n)$ in G **do**
    Extract Video Features:
    $\widetilde{\mathcal{X}}_v \leftarrow \mathcal{F}_d(\mathcal{X}_v^n)$ ▷ *Temporal downsampling via Conv1D*
    $\mathcal{X}_\mathcal{E} \leftarrow \mathcal{F}_{emb}(\widetilde{\mathcal{X}}_v)$ ▷ *Linear patch embedding*
    $\mathcal{X}_v^{(1)} \leftarrow (\mathcal{T}_{cls} \oplus \mathcal{X}_\mathcal{E}) + \mathcal{P}$ ▷ *Add CLS token and positional encoding*
    $\mathcal{X}_v^{(o)} \leftarrow \text{ViT-Residual}(\mathcal{X}_v^{(1)})$ ▷ *$\mathcal{N}$ transformer blocks with residual connections*
    Extract Audio Features:
    $\mathcal{X}'_f \leftarrow \mathcal{F}_{lnp}(\mathcal{X}_a^n)$ ▷ *Time-frequency projection and downsampling*
    $\mathcal{X}_p \leftarrow \text{PatchEmbed}(\mathcal{X}'_f)$ ▷ *2D convolutional patch embedding*
    $\mathcal{X}_{pe} \leftarrow [x_{cls}, x_{dist}, \mathcal{X}_p] + \mathcal{X}_e$ ▷ *Add CLS, distillation tokens, positional embedding*
    $\mathcal{X}_a^{(o)} \leftarrow \text{TransformerEncoder}(\mathcal{X}_{pe})$ ▷ *Multi-layer MHSA with layer normalization*
    Fuse Multimodal Features:
    if $S = \text{LT}$ **then** ▷ *Late Transformer Fusion*
        $\mathcal{O}_{V\to A} \leftarrow \text{CrossAttn}(\mathcal{X}_a^{(o)}, \mathcal{X}_v^{(o)})$ ▷ *Visual info into audio*
        $\mathcal{O}_{A\to V} \leftarrow \text{CrossAttn}(\mathcal{X}_v^{(o)}, \mathcal{X}_a^{(o)})$ ▷ *Audio info into video*
    else if $S = \text{IT}$ **then** ▷ *Intermediate Transformer Fusion*
        $\hat{\mathcal{X}}_a, \hat{\mathcal{X}}_v \leftarrow \text{Conv1D}(\mathcal{X}_a^{(o)}), \text{Conv1D}(\mathcal{X}_v^{(o)})$
        $\mathcal{O}_{V\to A} \leftarrow \text{CrossAttn}(\hat{\mathcal{X}}_a, \hat{\mathcal{X}}_v)$
        $\mathcal{O}_{A\to V} \leftarrow \text{CrossAttn}(\hat{\mathcal{X}}_v, \hat{\mathcal{X}}_a)$
    else if $S = \text{IA}$ **then** ▷ *Intermediate Attention Fusion*
        $\hat{\mathcal{X}}_a, \hat{\mathcal{X}}_v \leftarrow \text{Conv1D}(\mathcal{X}_a^{(o)}), \text{Conv1D}(\mathcal{X}_v^{(o)})$
        $\mathcal{O}_{V\to A} \leftarrow \text{DotProductAttn}(\hat{\mathcal{X}}_a, \hat{\mathcal{X}}_v)$
        $\mathcal{O}_{A\to V} \leftarrow \text{DotProductAttn}(\hat{\mathcal{X}}_v, \hat{\mathcal{X}}_a)$
    end if
    $Z \leftarrow \text{Pool-Concat}(\mathcal{O}_{V\to A}, \mathcal{O}_{A\to V})$
    Detect Depression:
    $p \leftarrow \text{Classifier}(Z)$
    if $p > 0.5$ **then**
        $c_n \leftarrow \text{Depression}$
    else
        $c_n \leftarrow \text{Normal}$
    end if
end for
return C

3.2.1 Video Feature Extraction

In the video feature extraction (as shown in top part of Fig. 3.5), the video data is first pre-processed and then embedded in a high-dimensional space suitable for transformer-based processing. The module utilizes a pre-trained ViT architecture [22] to process the video signals.

Initially, we downsample the video input using a 1D convolution block along the temporal dimension, where $\mathcal{T}$ is the sequence length, and $\mathcal{C}$ is the feature dimension for the input tensor $\mathcal{X}_v \in \mathbb{R}^{\mathcal{T}\times\mathcal{C}}$. This operation refines the input resolution to a fixed length $\mathcal{L}$ as mentioned in (3.18).

$$\widetilde{\mathcal{X}_v} = \mathcal{F}_d(\mathcal{X}_v) \tag{3.18}$$

where $\mathcal{F}_d(\cdot)$ represents the sequence of convolution, normalization, and pooling operations. Then, we apply a linear patch embedding ($\mathcal{F}_{emb}$) as in (3.19).

$$\mathcal{X}_{\mathcal{E}} = \mathcal{F}_{emb}(\widetilde{\mathcal{X}}_v) = \mathcal{W}_{emb}\widetilde{\mathcal{X}}_v \quad \in \mathbb{R}^{\mathcal{L}\times D} \tag{3.19}$$

where $\mathcal{W}_{emb} \in \mathbb{R}^{\mathcal{C}\times D}$ and $\mathcal{D}$ are the embedding weight and dimension.

After obtaining the patch embeddings $\mathcal{X}_{\mathcal{E}}$, a learnable classification token ($\mathcal{T}_{cls}$) is added at the beginning of the embedding sequence as shown in (3.20).

$$\mathcal{X}_v^{(0)} = \mathcal{T}_{cls} \oplus \mathcal{X}_{\mathcal{E}} \quad \in \mathbb{R}^{(\mathcal{L}+1)\times D} \tag{3.20}$$

In addition, learnable positional encoding $\mathcal{P} \in \mathbb{R}^{1\times(\mathcal{L}+1)\times D}$ is incorporated to capture spatial information as in (3.21).

$$\mathcal{X}_v^{(1)} = \mathcal{X}_v^{(0)} + \mathcal{P} \tag{3.21}$$

The token-augmented sequence $\mathcal{X}_v^{(1)}$ is subsequently passed through $\mathcal{N}$ transformer blocks ($n = 1, \ldots, \mathcal{N}$). In each block, self-attention ($\mathcal{Z}^n$) is computed as described in (3.22), where the queries, keys, and values are represented as $\mathcal{Q}^n = \mathcal{X}_v^{(n)}\mathcal{W}^{\mathcal{Q}}$, $\mathcal{K}^n = \mathcal{X}_v^{(n)}\mathcal{W}^{\mathcal{K}}$, and $\mathcal{V}^n = \mathcal{X}_v^{(n)}\mathcal{W}^{\mathcal{V}}$, respectively.

$$\mathcal{Z}^n = \mathcal{F}_{soft}\left(\frac{\mathcal{Q}^n(\mathcal{K}^n)^\top}{\sqrt{d}}\right)\mathcal{V}^n \tag{3.22}$$

here $\mathcal{W}^{\mathcal{Q}}$, $\mathcal{W}^{\mathcal{K}}$, $\mathcal{W}^{\mathcal{V}} \in \mathbb{R}^{D\times D}$ are the learnable weight matrices, $\mathcal{F}_{soft}$ is the softmax activation, and d is the dimension for each head.

A residual connection followed by a MLP with layer normalization is applied to the output of each transformer block as shown in (3.23).

$$\mathcal{X}_v^{(n+1)} = \mathcal{F}_{mlp}(\mathcal{Z}^n + \mathcal{X}_v^{(n)}) + \left(\mathcal{Z}^n + \mathcal{X}_v^{(n)}\right) \tag{3.23}$$

This mechanism is applied repeatedly, where the output of the final transformer block, $\mathcal{X}_v^{(o)} \in \mathbb{R}^{(\mathcal{L}+1)\times D}$ serves as the high-level visual feature representation from the video input.

3.2.2 Audio Feature Extraction

The module processes an input audio waveform $\mathcal{X}_a \in \mathbb{R}^S$ (S denotes the number of samples) through a series of transformations, including linear projection, patch and positional embedding, and transformer encoding (as illustrated in middle part of Fig. 3.5).

Initially, the waveform is transformed into a time-frequency representation $\mathcal{X}_f \in \mathbb{R}^{\mathcal{F}\times\mathcal{T}}$, where $\mathcal{F}$ and $\mathcal{T}$ represent the number of frequency bins and time frames. To ensure consistent input dimensions, $\mathcal{X}_f$ is projected to a fixed-size matrix $\mathcal{X}'_f \in \mathbb{R}^{\mathcal{F}'\times\mathcal{T}'}$ via a learnable linear projection function $\mathcal{F}_{lnp}$ as in (3.24).

$$\mathcal{X}'_f = \mathcal{F}_{lnp}(\mathcal{X}_f; \mathcal{F}', \mathcal{T}', \theta_d) \tag{3.24}$$

where θ_d represents the parameters of convolutional and pooling operations. In particular, a 1D convolution adjusts the frequency dimension, followed by batch normalization and adaptive average pooling to standardize the time dimension to $\mathcal{T}'$. The resulting matrix is then reshaped into a single-channel 2D input $\mathcal{X}'_f \in \mathbb{R}^{1\times\mathcal{F}'\times\mathcal{T}'}$.

The obtained feature matrix $\mathcal{X}'_f$ is partitioned into overlapping patches using a 2D convolutional layer to generate patch embeddings $\mathcal{X}_p \in \mathbb{R}^{\mathcal{M}\times D}$. The convolution operates with patch sizes (p_f, p_t) and strides (s_f, s_t), generating a feature map with spatial dimensions $h = \lfloor(\mathcal{F}' - p_f)/s_f\rfloor + 1$ and $w = \lfloor(\mathcal{T}' - p_t)/s_t\rfloor + 1$, resulting in $\mathcal{M} = h \times w$ patches. The resulting feature map in 3D tensor form is flattened and transposed using the operation $\mathcal{F}_{flt}(\cdot)$ as in (3.25), where $\tilde{\mathcal{X}}_p$ denotes the intermediate output of the convolution that makes the sequence appropriate for transformer encoder.

$$\mathcal{X}_p = \mathcal{F}_{flt}(\tilde{\mathcal{X}}_p) \in \mathbb{R}^{\mathcal{M}\times D} \tag{3.25}$$

A base positional embedding matrix $\mathcal{X}_{e_{base}} \in \mathbb{R}^{D\times h_{base}\times w_{base}}$, following the AST [23], is resized using bilinear interpolation to align with the patch grid size (h, w) as mentioned in (3.26).

$$\mathcal{X}_e = \mathcal{F}_{inp}(\mathcal{X}_{e_{base}}; (h, w)) \in \mathbb{R}^{\mathcal{M}\times D} \tag{3.26}$$

Moreover, two special tokens are added to the sequence: a classification token x_{cls} and a distillation token x_{dist}, each with fixed positional embeddings e_{cls} and e_{dist}. The final embedded sequence is generated as in (3.27).

$$\begin{aligned}\mathcal{X}_{pe} = [&x_{cls} + e_{cls},\ x_{dist} + e_{dist},\\ &\mathcal{X}_{p_1} + \mathcal{X}_{e_1},\ \ldots,\ \mathcal{X}_{p_\mathcal{M}} + \mathcal{X}_{e_\mathcal{M}}] \in \mathbb{R}^{(\mathcal{M}+2)\times D}\end{aligned} \tag{3.27}$$

The embedded sequence $\mathcal{X}_{pe}$, obtained from (3.27), is processed by a transformer encoder comprising several identical layers. Each layer consists of a multi-head self-attention ($\mathcal{F}_{mhsa}$)

followed by a feed-forward network. Each sub-layer is preceded by layer normalization ($\mathcal{F}_{ln}$) and followed by a residual connection. The computations within a single transformer layer are shown in (3.28).

$$\begin{aligned} \mathcal{U} &= \mathcal{F}_{ln}\left(\mathcal{X}_{pe} + \mathcal{F}_{mhsa}(\mathcal{X}_{pe})\right), \\ \mathcal{Z} &= \mathcal{F}_{ln}\left(\mathcal{U} + \mathcal{F}_{mlp}(\mathcal{U})\right) \end{aligned} \tag{3.28}$$

where $\mathcal{Z}$ corresponds to the final acoustic output sequence $\mathcal{X}_a^{(o)} \in \mathbb{R}^{(\mathcal{M}+2)\times D}$, capturing temporal feature representations.

3.2.3 Fusion Module

The fusion module in MMFformer combines extracted visual and acoustic features using three distinct strategies, each operating at a different level of the processing pipeline. The proposed fusion module is described as illustrated in the fusion module of the bottom part of Fig. 3.5. The following subsections describe each fusion approach in detail.

3.2.3.1 Late Transformer Fusion

This architecture fuses the extracted visual and acoustic features using transformer blocks. Each network employs its own transformer block to perform cross-modal fusion. The video network takes visual features $\mathcal{X}_v^{(o)}$, while the audio network processes acoustic features $\mathcal{X}_a^{(o)}$, and fusion happens by combining features from one modality into the other. Each network processes its respective features through a series of 1D Convolution (Conv1D) layers. The outputs of each transformer block are then pooled, concatenated and fed into a final depression detection layer.

For the acoustic network, the transformer block takes the visual network representation $\mathcal{X}_v^{(o)}$ as input to compute the keys and values, while queries are derived from the features of the acoustic network $\mathcal{X}_a^{(o)}$. The self-attention mechanism is computed in (3.29).

$$\mathcal{O} = \mathcal{F}_{soft}\left(\frac{\mathcal{X}_a^{(o)}\mathcal{W}_q\mathcal{W}_k^{\top}(\mathcal{X}_v^{(o)})^{\top}}{\sqrt{d}}\right)\mathcal{X}_v^{(o)}\mathcal{W}_v \tag{3.29}$$

where $\mathcal{W}_q$, $\mathcal{W}_k$, and $\mathcal{W}_v$ are the weight matrices for queries, keys, and values, respectively. This operation fuses visual information into the acoustic network ($V \rightarrow A$). Similarly, the visual network transformer block computes queries from $\mathcal{X}_v^{(o)}$, and keys and values from $\mathcal{X}_a^{(o)}$, allowing fusion of acoustic information into the visual network ($A \rightarrow V$).

3.2.3.2 Intermediate Transformer Fusion

This module introduces cross-modal fusion at an intermediate stage, enabling earlier interaction between visual and acoustic features. The extracted features are input into two Conv1D layers in each branch to obtain the intermediate representations $\mathcal{X}_v^{(o)}$ and $\mathcal{X}_a^{(o)}$, which are passed through separate transformer blocks for cross-modal fusion as shown in (3.29).

In the acoustic network, queries are computed from $\mathcal{X}_a^{(o)}$, while keys and values are derived from $\mathcal{X}_v^{(o)}$, enabling visual-to-acoustic information transfer. However, the visual network receives acoustic features through a symmetric operation. The outputs of each network are then fused separately and passed through an additional Conv1D layer. The refined features then pass through the pooling, concatenation, and final classification layer for depression detection.

3.2.3.3 Intermediate Attention Fusion

This architecture presents an attention-based fusion at an intermediate level, enabling cross-modal interaction without directly fusing feature representations. The extracted features are fed into two Conv1D layers separately, resulting in visual and acoustic features $\mathcal{X}_v^{(o)}$ and $\mathcal{X}_a^{(o)}$, which are processed through attention mechanisms by exploiting dot-product similarity to highlight mutually relevant features.

In the acoustic network, queries are computed from $\mathcal{X}_a^{(o)}$ and keys from $\mathcal{X}_v^{(o)}$. The scaled dot-product attention is calculated in (3.30).

$$\mathcal{O} = \mathcal{F}_{soft}\left(\frac{\mathcal{X}_a^{(o)}\mathcal{W}_q\mathcal{W}_k^{\top}(\mathcal{X}_v^{(o)})^{\top}}{\sqrt{d}}\right) \tag{3.30}$$

The softmax operation emphasizes the most salient features of the visual modality relative to the acoustic features ($V \rightarrow A$). The visual network performs the same mechanism to compute attention from acoustic features ($A \rightarrow V$).

The attention vector for the visual network is calculated as $v_v = \sum_{i=1}^{N_v} \mathcal{O}[:, i]$, capturing the most relevant visual attributes based on their alignment with acoustic features. The acoustic network follows a symmetric process. The attention-weighted outputs are refined through an additional Conv1D layer, pooled, concatenated, and passed to the depression detection layer.

3.3 Summary

This chapter presented the detailed architectures of the two proposed frameworks for multimodal depression detection from social media video blogs. MDD-Net addresses the cross-modal fusion problem through a MT that bidirectionally computes audio-to-video and video-to-audio correlations, complemented by a joint transformer that further refines the concatenated representations. The architecture operates on hand-crafted features, specifically 25 OpenSMILE extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) descriptors for audio and 68 Dlib facial landmarks for video, and employs a composite loss function combining binary cross-entropy, focal loss, and L2 regularization to handle the noisy labels and class imbalance characteristic of social media datasets. MMFformer takes a different approach by employing pre-trained backbone networks, a residual ViT for video and an audio spectrogram transformer for audio, and systematically comparing three fusion strategies that operate at different levels of the processing pipeline. The late transformer fusion allows cross-modal interaction only after each branch has fully processed its input, the intermediate transformer fusion introduces interaction at an earlier stage, and the intermediate attention fusion uses dot-product attention to highlight mutually relevant features without directly merging the representations.

Both architectures share the core design principle that audio and visual modalities should interact via cross-attention mechanisms rather than through symmetric operations such as concatenation or element-wise addition. This principle aligns with the broader trend in multimodal AI, where recent VLMs and MLLMs such as NExT-GPT [55], Unified-IO 2 [56], and Show-o [57] have demonstrated that cross-modal attention within unified transformer frameworks can capture richer interactions across modalities than traditional fusion pipelines. While these large-scale models operate at fundamentally different scales and are designed for general-purpose multimodal understanding, the cross-attention mechanisms explored in MDD-Net and MMFformer address the same underlying challenge: learning how information from one modality can inform and modulate the representation of another. The experimental evaluation of both frameworks is presented in the following chapter.

Chapter 4

Experiments and Results

This chapter presents the experimental evaluation of both proposed frameworks. It begins with a description of the datasets used, followed by the implementation details and evaluation metrics. The chapter then reports the performance of MDD-Net and MMFformer, including comparisons with SOTA methods, ablation studies, visual analysis, and cross-corpus validation experiments.

4.1 Datasets

4.1.1 D-Vlog Dataset

The D-Vlog dataset [16], collected from YouTube, consists of 961 vlog videos spanning approximately 160 hours of recordings from 816 individuals (322 men and 639 women). It includes 555 depressed and 406 normal samples. The depressed class labels are assigned based on specific keywords in the titles, with terms like "depression episode vlog", "depression daily vlog", "my depression story", "depression journey", "my depression diary", "depression video diary", and "depression vlog". Moreover, the non-depressed labels contain "haul vlog", "daily vlog", "grwm (get ready with me) vlog", "how to vlog", "talking vlog", and "day of vlog". The authors carried out two tasks to ensure the labels were valid: verifying that each video adhered to the vlog format, with individuals speaking directly to the camera, and employing annotators to analyze videos for signs of depression using automatically generated transcripts. The database is partitioned into training, validation, and test sets with a ratio of 7:1:2. Acoustic features are obtained using OpenSMILE with eGeMAPS, yielding 25 low-level descriptors per second that capture properties such as loudness, pitch, spectral flux, and MFCC. The visual features consist of

68 facial landmarks per frame, extracted using the Dlib library and encoded as 136-dimensional vectors representing the x and y coordinates of each landmark. These features are sampled at a per-second interval, making them appropriate for temporal analysis. D-Vlog is unique in collecting real-world, unscripted videos in which individuals openly share their daily lives and mental health.

4.1.2 LMVD Dataset

The LMVD [17] is a large-scale dataset for detecting depression in everyday contexts, used in this thesis to evaluate the performance of MMFformer. The dataset comprises 1,823 vlog samples totaling around 214 hours of video, collected from 1,475 people across four platforms: Bilibili, TikTok, Sina Weibo, and YouTube. Each video is classified as either depressed or non-depressed, following a manual assessment by volunteers and validation by clinical professionals. The audio features are 128-dimensional embeddings derived from the VGGish network, which operates on log-mel spectrograms rather than hand-crafted acoustic descriptors. The visual features are more comprehensive than those in D-Vlog, incorporating facial action units, face landmarks, eye gaze vectors, and head pose estimates extracted through a combination of Dlib and OpenFace. LMVD is a heterogeneous dataset of real-world significance, as the videos are spontaneous and self-recorded, accurately depicting real-user behavior rather than being generated in laboratory-controlled environments.

4.1.3 Comparison of Dataset Characteristics

Table 4.1 summarizes the key differences between the two benchmark datasets. Beyond the difference in scale, the two datasets employ fundamentally different feature extraction pipelines. D-Vlog relies on hand-crafted acoustic descriptors through the eGeMAPS feature set, while LMVD uses learned audio representations from VGGish that encode spectral patterns at a higher dimensionality (128 versus 25). On the visual side, D-Vlog provides only facial landmark coordinates, whereas LMVD offers a richer set of features, including facial action units, eye gaze, and head pose, alongside landmarks. The cross-platform nature of LMVD also introduces greater variability in recording conditions, video quality, and speaker demographics. These differences are important to consider because they directly affect the input dimensions and preprocessing requirements of the proposed architectures, and they make the cross-corpus validation experiments in Section 4.4.4 a meaningful test of generalizability.

Table 4.1: Dataset Feature Specifications for D-Vlog and LMVD.

Specification	**D-Vlog**	**LMVD**
Total Videos	961	1,823
Total Individuals	816	1,475
Total Duration	∼160 hours	∼214 hours
Source Platforms	YouTube	Bilibili, TikTok, Sina Weibo, YouTube
Classes	Depressed / Normal	Depressed / Normal
Data Split	7:1:2	7:1:2
Audio Features		
Extraction Tool	OpenSMILE	VGGish
Feature Set	eGeMAPS (25 LLDs)	Audio embeddings
Sampling Rate	Per-second	Per-second
Feature Dimension	25	128
Visual Features		
Extraction Tool	Dlib	Dlib + OpenFace
Feature Type	68 face landmarks	FAU, landmarks, gaze, head pose
Sampling Rate	Per-second	Per-second
Feature Dimension	136 (68×2)	Multiple vectors

4.2 Experimental Setup

4.2.1 MDD-Net Implementation Details

All experiments are implemented using the PyTorch framework in Python. The proposed architecture is executed on 48 GB NVIDIA RTX A6000 graphics card. The Adam optimizer is employed with the following settings: learning rate = 1e-4, weight decay = 0.1, and epsilon = 1e-8. The network is trained for 200 epochs with a batch size of 8. To prevent overfitting, an early stopping criterion is applied to stop training if no improvement in the validation error is observed for 15 epochs, saving the best model based on the validation set performance. A 10-fold cross-validation is implemented to ensure the reported results are not dependent on a single data partition.

4.2.2 MMFformer Implementation Details

The MMFformer is developed and trained using the PyTorch deep learning framework in Python. The batch size is set to 16 during training, and the maximum number of training epochs is 225. For optimization, Adam is selected as the optimizer with a learning rate = 1e-5, weight decay = 0.1, and epsilon = 1e-8. An early stopping mechanism with a patience of 15 epochs is also implemented to halt training when validation performance stops improving. The performance of the developed system is validated through a 10-fold cross-validation. The proposed architecture is evaluated on two 48 GB NVIDIA RTX A6000 GPUs.

4.2.3 Comparison of Training Configurations

Table 4.2 provides a side-by-side summary of the training hyperparameters for both frameworks. While both models share the same optimizer, regularization strategy, and early stopping protocol, several differences in their configurations reflect the distinct architectural requirements of each framework. MMFformer uses a learning rate ten times smaller than MDD-Net (1e-5 versus 1e-4) to avoid destabilizing the pre-trained ViT and AST weights during fine-tuning, since MDD-Net trains its feature extraction modules from scratch and benefits from a higher initial learning rate. The larger batch size in MMFformer (16 versus 8) helps stabilize gradient estimates when training across two datasets with different feature distributions, whereas MDD-Net operates on a single dataset and does not require this. MMFformer is also given 25 additional training epochs to allow its more complex fusion modules sufficient time to converge. The difference in GPU requirements follows from the same reasoning: the pre-trained backbone networks and multiple fusion strategy experiments in MMFformer introduce computational overhead that necessitates two GPUs, while MDD-Net fits comfortably on a single card. The values are obtained from the experimental setup described in the respective papers and confirmed through the publicly available source code repositories.[1] [2]

[1] MDD-Net: https://github.com/rezwanh001/Multimodal-Depression-Detection
[2] MMFformer: https://github.com/rezwanh001/Large-Scale-Multimodal-Depression-Detection

Table 4.2: Training Hyperparameters for MDD-Net and MMFformer.

Parameter	**MDD-Net**	**MMFformer**
Optimizer	Adam	Adam
Learning Rate	1e-4	1e-5
Weight Decay	0.1	0.1
Epsilon (ϵ)	1e-8	1e-8
Max Epochs	200	225
Batch Size	8	16
Early Stopping Patience	15 epochs	15 epochs
Cross-Validation	10-fold	10-fold
Framework	PyTorch	PyTorch
GPU(s)	1× RTX A6000	2× RTX A6000

4.2.4 Evaluation Metrics

Four widely recognized performance metrics such as Accuracy (Acc), Precision (Pr), Recall (Rc), and F1-Score (F1) are used to assess the performance. These metrics can evaluate overall performance on both balanced and unbalanced datasets. For MDD-Net, these metrics are computed using standard definitions since the D-Vlog data set is the only benchmark. For MMFformer, to better understand how well the model performs across different class distributions, we report both Weighted Average (WA) and Unweighted Average (UA) for each metric.

4.3 MDD-Net Results

4.3.1 Performance Analysis

The performance of the proposed MDD-Net for depression detection from multimodal vlog data is summarized in Table 4.3. The results demonstrate the advantage of multimodal fusion over unimodal approaches.

Table 4.3: Performances of Proposed MDD-Net for Depression Detection from Multimodal Vlog Data.

Methods	Features	Acc	Pr	Rc	F1
AFEM	Acoustic	0.4770	0.5552	0.4667	0.4106
VFEM	Visual	0.4604	0.5551	0.4577	0.4239
MDD-Net	Both	0.7260	0.7392	0.8065	0.7707

The MDD-Net, with multimodal data, outperforms the acoustic feature module by 24.90% in accuracy and 36.01% in F1-Score. Moreover, MDD-Net shows a 26.56% enhancement in accuracy and a 34.68% increase in F1-Score compared to the visual feature block. The found results highlight the value of using mutual transformers to fuse features before passing them to the detection layer. The proposed fusion approach in the MDD-Net architecture has been shown to improve classifier performance, making it more reliable at distinguishing between normal and depression cases across different data modalities.

4.3.2 Comparison with the State-of-the-Art

In this section, the performance of MDD-Net is compared with other methods from the literature, as shown in Table 4.4. Experimental results show that the proposed MDD-Net improves recall and F1-score by fusing features extracted from both the audio and visual modules using mutual transformers. The percentage improvement to literature for deep learning methods, such as Bi-LSTM [58] and TFN [11], ranges from 59.70% to 61.00% in F1-Score. More recent approaches focus on advanced feature fusion to further enhance results. For instance, the Depression Detector [16] combines audio and visual features via cross-transformers, while Time-Aware Attention Multimodal Fusion Network (TAMFN) [32] applies time-aware attention to integrate multimodal features. CAIINET [37] utilizes contextual attention and interaction to capture significant temporal information and fuse multimodal features. STST [45] adds spatio-temporal information from various modalities for better results. Similarly, MDAVIF [36] fuses features across modalities and leverages two autoencoders to preserve information and improve performance. The MDD-Net, proposed in this work, achieves higher performances with an enhancement of F1-Score ranging from 1.82% to 17.37%, with a precision of 0.7392, and a recall of 0.8065, showing robust performance in detecting depression.

Table 4.4: Performance Comparison of the Proposed MDD-Net with Different Existing Methods on D-Vlog Dataset.

Methods	Pr	Rc	F1
Bi-LSTM [58]	0.6081	0.6179	0.5970
TFN [11]	0.6139	0.6226	0.6100
Depression Detector [16]	0.6540	0.6557	0.6350
TAMFN [32]	0.6602	0.6650	0.6582
CAIINET [37]	0.6656	0.6698	0.6655
STST [45]	0.7250	0.7767	0.7500
MDAVIF [36]	**0.7425**	0.7600	0.7525
MDD-Net	0.7392	**0.8065**	**0.7707**

4.3.3 Ablation Study

To test the effectiveness of each module in MDD-Net, ablation experiments are carried out, and their outcomes are analyzed in this section. To assess the impact of various fusion methods on the performance of the MDD-Net model, a series of ablation studies are performed and their findings are reported in Table 4.5. The results show that the MT fusion method outperforms the other fusion techniques by up to 17.37% for F1-Score, achieving the highest accuracy of 0.7260, precision of 0.7392, and recall of 0.8065. In comparison, the Addition (Add) fusion method achieves an accuracy of 0.6927, with precision, recall, and F1-Score of 0.7086, 0.7909, and 0.7445, respectively. Similarly, the Concatenation (Concat) technique achieves promising outcomes with an accuracy of 0.6979 and F1-Score of 0.7469, while the Multiplication (Multi) approach has a 4.06% and 3.69% lower performance across the accuracy and F1-Score. These findings highlight the efficiency of the MT block in capturing and fusing the interactions between acoustic and visual features, thus enhancing the network's overall performance in multimodal depression recognition. Hence, the ablation studies reveal that the design of the MT module plays a crucial role in utilizing diverse feature sets for improved classification outcomes.

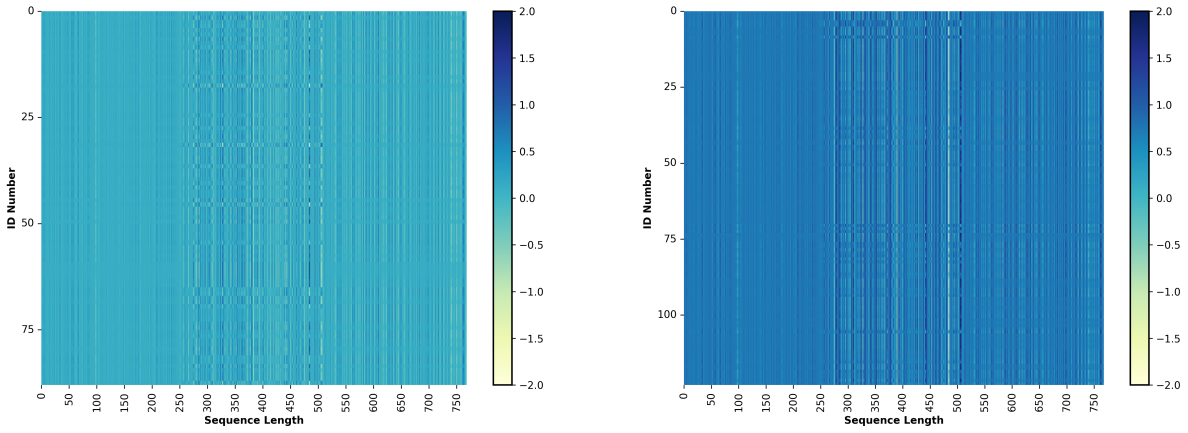
Table 4.5: Results of Ablation Studies for the Proposed MDD-Net Architecture to Detect Depression.

Fusion	Acc	Pr	Rc	F1
Add	0.6927	0.7086	0.7909	0.7445
Multi	0.6854	0.7125	0.7721	0.7338
Concat	0.6979	0.7214	0.7799	0.7469
MDD-Net	**0.7260**	**0.7392**	**0.8065**	**0.7707**

4.3.4 Visual Results

In this section, feature maps are presented after fusing audio and video modalities. The t-Distributed Stochastic Neighbor Embedding (t-SNE) feature visualization is utilized to provide the qualitative results [59].

Figure 4.1 shows the weight visualizations for both classes based on the fused acoustic-visual features. The weights are visualized on the final embeddings, which combine acoustic and visual features. Using the trained network, the weights are visualized using test samples, and the differences in sequence length between depressed and non-depressed individuals are analyzed. In



(a) People with Normal

(b) People with Depression

Figure 4.1: Visualization of features based on weights of the final embedding obtained by fusing audio-visual features; the darker the color, the higher the weights.

Fig. 4.1, the darker blue colors indicate higher weights for the depressed class, while lighter blue colors refer to lower weights for the normal class.

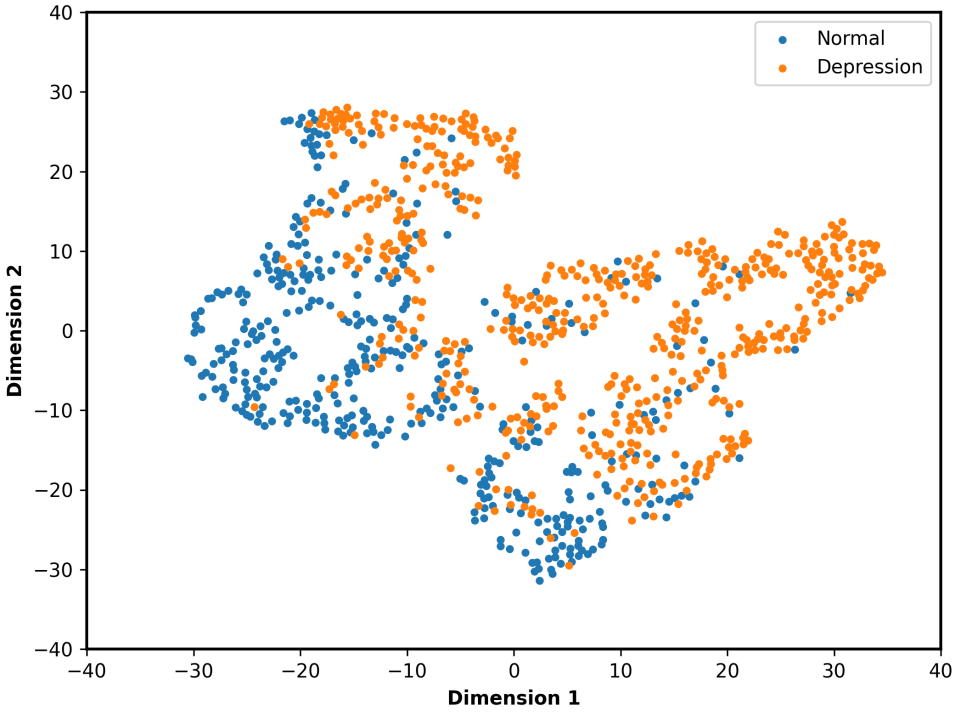


Figure 4.2: Feature visualization of the proposed MDD-Net through t-SNE for multimodal depression detection.

Figure 4.2 illustrates the distribution of fused acoustic-visual features from the D-Vlog dataset using t-SNE visualization. Some normal samples overlap with the depressed category, leading to misclassifications. These misclassifications occur because the MDD-Net struggles to distinguish these categories but can still detect clusters.

4.4 MMFformer Results

This section presents an extensive set of experiments to assess the performance of the proposed network for detecting depression. We begin by exploring diverse combinations of features derived

using video and audio networks and different methods for fusing these features. From these preliminary results, the most promising designs are selected and their performance is evaluated against several existing models. Ablation studies are performed to gain insight into the distinct effects of various fusion methods concerning the proposed architecture. Lastly, cross-corpus experiments are conducted to evaluate generalizability across multiple datasets.

4.4.1 Depression Detection Results

Table 4.6 presents the performance of MMFformer for depression detection on the D-Vlog and LMVD datasets. The results are reported using Video (V), Audio (A), and Audio+Video (A+V) modalities with three fusion strategies: LT, IT, and IA. Each experiment was repeated 10 times, and the mean and standard deviation (mean ± std) were reported for all performance metrics.

For both datasets, multimodal fusion outperforms unimodal approaches. On the D-Vlog dataset, the IT fusion architecture achieves the highest Weighted Average Accuracy (WAA) of 0.8108, Weighted Average Precision (WAP) of 0.8924, and Weighted Average F1-Score (WAF1) of 0.9092, while the IA fusion receives the highest Weighted Average Recall (WAR) of 0.9380. Similarly, IT network obtains Unweighted Average Accuracy (UAA) of 0.7957, Unweighted Average Precision (UAP) of 0.9088, and Unweighted Average F1-Score (UAF1) of 0.9239 for unweighted metrics, while IA fusion performs the best in Unweighted Average Recall (UAR) of 0.9471. On the LMVD dataset, the LT model scores the highest performance across all weighted and unweighted metrics, including WAA of 0.8071, WAP of 0.9013, WAR of 0.9112, and WAF1 of 0.9048. The IT and IA models perform closely, with IA slightly outperforming IT in WAP and UAA.

Overall, the results highlight the enhanced performance of fusion-based multimodal learning

Table 4.6: Performances of MMFformer for Depression Detection from Multimodal Vlog Data. Each Model is Run Ten Times to Obtain the Results (Mean ± Std). Bold Represents the Best and Underline Indicates the Second-best.

Dataset	Modalities	Fusion	WAA	WAP	WAR	WAF1	UAA	UAP	UAR	UAF1
D-Vlog	A	–	0.7814 ± 0.039	0.8786 ± 0.030	0.9165 ± 0.031	0.8955 ± 0.018	0.7638 ± 0.044	0.8969 ± 0.028	0.9297 ± 0.028	0.9115 ± 0.018
	V	–	0.7214 ± 0.051	0.8438 ± 0.031	0.9041 ± 0.037	0.8704 ± 0.021	0.6912 ± 0.064	0.8682 ± 0.028	0.9188 ± 0.034	0.8904 ± 0.021
	A+V	LT	0.7958 ± 0.031	0.8779 ± 0.027	0.9367 ± 0.020	0.9046 ± 0.012	0.7731 ± 0.042	0.8966 ± 0.026	**0.9473 ± 0.016**	0.9196 ± 0.011
		IT	**0.8108 ± 0.026**	**0.8924 ± 0.024**	0.9308 ± 0.037	**0.9092 ± 0.014**	**0.7957 ± 0.029**	**0.9088 ± 0.024**	0.9432 ± 0.028	**0.9239 ± 0.009**
		IA	0.8030 ± 0.029	0.8806 ± 0.019	**0.9380 ± 0.026**	0.9071 ± 0.014	0.7776 ± 0.037	0.8995 ± 0.017	0.9471 ± 0.024	0.9215 ± 0.014
LMVD	A	–	0.7175 ± 0.024	0.8621 ± 0.023	0.8563 ± 0.030	0.8575 ± 0.011	0.7143 ± 0.024	0.8572 ± 0.024	0.8502 ± 0.036	0.8519 ± 0.018
	V	–	0.6905 ± 0.055	0.8452 ± 0.022	0.8515 ± 0.074	0.8437 ± 0.043	0.6914 ± 0.053	0.8399 ± 0.023	0.8478 ± 0.070	0.8393 ± 0.039
	A+V	LT	**0.8071 ± 0.022**	**0.9013 ± 0.018**	**0.9112 ± 0.034**	**0.9048 ± 0.013**	**0.8089 ± 0.022**	**0.8966 ± 0.025**	**0.9090 ± 0.031**	**0.9014 ± 0.016**
		IT	0.8035 ± 0.029	0.8994 ± 0.018	0.9072 ± 0.028	0.9024 ± 0.016	0.8019 ± 0.031	0.8955 ± 0.022	0.9031 ± 0.033	0.8984 ± 0.021
		IA	0.8023 ± 0.022	0.9006 ± 0.020	0.9064 ± 0.034	0.9019 ± 0.013	0.8032 ± 0.022	0.8960 ± 0.025	0.9036 ± 0.033	0.8983 ± 0.016

methods over unimodal baselines. Among fusion strategies, IT has the best performance with the D-Vlog dataset, while LT has the highest performance with the LMVD dataset. The findings indicate that dataset features affect the optimal fusion method, highlighting the significance of customized architecture for multimodal depression detection.

4.4.2 Comparison with the State-of-the-Art

To evaluate the efficiency of our proposed architecture, we compared its performance with several SOTA approaches in the D-Vlog and LMVD datasets (as summarized in Table 4.7). On the LMVD dataset, our model achieves an F1-Score of 0.9048 and a precision of 0.9013, outperforming all existing methods, including the work presented in [42]. That work reports the second-best F1-Score of 0.8274 and precision of 0.8287, achieving relative improvements of 7.74% in F1-Score and 7.26% in precision. Although the system developed in [42] achieved the highest accuracy of 0.8276, our approach demonstrates superior precision of 0.9013 and recall of 0.9112, indicating better reliability for depression detection. On the D-Vlog dataset, MMFformer outperforms existing methods across all evaluation metrics. An accuracy of 0.8108, a precision of 0.8924, a recall of 0.9380, and an F1-Score of 0.9092 are obtained; all of which surpass previously reported results. The proposed system achieves a relative increase of 5.54% in accuracy, 13.92% in F1-Score, and 6.81% in recall compared to prior methods. The system developed in [45] achieved an F1-Score of 0.7500, with noticeably lower accuracy and precision. While some methods, such as [38] and [41] reported competitive recall and F1-Scores, their precision values were comparatively low.

Table 4.7: Comparison with the SOTA Methods on D-Vlog and LMVD Datasets for Depression Detection.

Methods	Datasets	Acc	Pr	Rc	F1
Ye et al. [38]	LMVD	0.7213	0.7018	0.7656	0.7320
He et al. [42]		**0.8276**	0.8287	0.8276	0.8274
Ye et al. [38]	D-Vlog	0.6887	0.6819	0.8699	0.7644
Xing et al. [39]		0.7554	0.7579	0.7554	0.7566
Shangguan et al. [40]		–	0.6727	0.6777	0.6664
Zhou et al. [37]		–	0.6656	0.6698	0.6655
Kowalewski et al. [41]		–	0.7100	0.8400	0.7700
He et al. [43]		0.6500	0.6400	0.5400	0.5900
Tao et al. [44]		–	0.7153	0.7560	0.7351
Tao et al. [45]		0.7070	0.7250	0.7767	0.7500
Yang et al. [46]		0.7073	–	–	–
MMFformer	D-Vlog	**0.8108**	**0.8924**	**0.9380**	**0.9092**
	LMVD	0.8071	**0.9013**	**0.9112**	**0.9048**

4.4.3 Ablation Study

To thoroughly assess the contributions of commonly used fusion strategies in MMFformer for depression detection, we conducted ablation studies on the D-Vlog and LMVD datasets. The outcomes of the ablation studies are presented in Table 4.8. The ablation experiments focused on combining audio and video modalities using four fusion methods: Add, Multi, Concat, and Tensor Fusion (TF) network [11]. The Concat achieved the best performance on the D-Vlog dataset, scoring WAA of 0.7652, WAP of 0.8794, WAF1 of 0.8833, and UAF1 of 0.9016. The Add method ranked second-best, with a high WAR of 0.9112 and WAF1 of 0.8860, demonstrating its ability to recognize depressive samples. The Add fusion performed best on the LMVD dataset, with a WAA of 0.7930, WAP of 0.8891, WAR of 0.9120, and WAF1 of 0.8998. It also scored the highest UAF1 of 0.8961. The Concat was the second-best on LMVD, with WAP of 0.8937 and WAF1 of 0.8833, illustrating its consistent performance across datasets. Considering TF fusion, it performed worst on both datasets, with WAF1 scores of 0.7357 and 0.8680 on the D-Vlog and LMVD, respectively, highlighting its weakness in effectively fusing multi-modal features. In a similar way, the Multi fusion underperformed, achieving WAF1 values of 0.8333 for D-Vlog and 0.8813 for LMVD. These findings revealed that the Concat and Add fusions are better at capturing and fusing complementary information from audio and video modalities. However, our proposed

Table 4.8: Results of Ablation Studies for MMFformer on the Video and Audio Data to Detect Depression.

Dataset	Modalities	Fusion	WAA	WAP	WAR	WAF1	UAA	UAP	UAR	UAF1
D-Vlog	A+V	Add	0.7606 ± 0.032	0.8658 ± 0.024	**0.9112 ± 0.038**	**0.8860 ± 0.015**	0.7385 ± 0.038	0.8866 ± 0.023	**0.9255 ± 0.033**	**0.9039 ± 0.015**
		Multi	0.7045 ± 0.059	0.8032 ± 0.124	0.8813 ± 0.091	0.8333 ± 0.099	0.6605 ± 0.095	0.8123 ± 0.163	0.8756 ± 0.136	0.8365 ± 0.146
		Concat	**0.7652 ± 0.033**	**0.8794 ± 0.017**	0.8890 ± 0.035	0.8833 ± 0.019	**0.7512 ± 0.035**	**0.8985 ± 0.015**	0.9063 ± 0.032	0.9016 ± 0.018
		TF	0.6734 ± 0.095	0.6796 ± 0.241	0.8241 ± 0.175	0.7357 ± 0.218	0.6189 ± 0.115	0.6893 ± 0.270	0.8132 ± 0.207	0.7371 ± 0.248
LMVD	A+V	Add	**0.7930 ± 0.029**	0.8891 ± 0.017	**0.9120 ± 0.021**	**0.8998 ± 0.014**	**0.7937 ± 0.029**	0.8846 ± 0.021	**0.9092 ± 0.021**	**0.8961 ± 0.016**
		Multi	0.7615 ± 0.039	0.8809 ± 0.024	0.8846 ± 0.036	0.8813 ± 0.020	0.7601 ± 0.037	0.8760 ± 0.030	0.8797 ± 0.040	0.8763 ± 0.026
		Concat	0.7724 ± 0.025	**0.8937 ± 0.018**	0.8745 ± 0.019	0.8833 ± 0.012	0.7702 ± 0.025	**0.8897 ± 0.022**	0.8698 ± 0.023	0.8789 ± 0.016
		TF	0.7252 ± 0.055	0.8567 ± 0.033	0.8849 ± 0.044	0.8680 ± 0.026	0.7219 ± 0.053	0.8522 ± 0.033	0.8792 ± 0.050	0.8628 ± 0.030

architecture exceeds all these outcomes, where the IT fusion achieves WAA of 0.8108 and WAF1 of 0.9092 on D-Vlog, and the LT method on LMVD records WAA of 0.8071 and WAF1 of 0.9048, demonstrating superior and consistent performance in detecting depression.

4.4.4 Cross-Corpus Validation

To evaluate the generalizability of MMFformer across different datasets, a cross-corpus validation is conducted between D-Vlog and LMVD, focusing on multimodal features. The results of cross-corpus experiments are shown in Table 4.9. Three fusion methods: LT, IT, and IA were tested in two experimental setups: (i) training on D-Vlog and testing on LMVD, and (ii) training on LMVD and testing on D-Vlog. In the first case, when trained on D-Vlog and tested on LMVD, the IA fusion achieved the highest performance, with WAA of 0.7454, WAP of 0.8784, and WAF1 of 0.8715. It also recorded a UAF1 of 0.8660, while the IT method reported a WAA of 0.7170. In the second case, when trained on LMVD and tested on D-Vlog, the IA method again performed well, achieving WAF1 of 0.8562 and high WAR of 0.9172, while the IT fusion recorded WAF1 of 0.8561 and UAF1 of 0.8781. The LT method shows the lowest performance in both scenarios, with WAF1 scores of 0.8529 and 0.8188, respectively. These results indicate that the IA fusion ensured generalizability across multiple datasets, due to its ability to capture and fuse multimodal

Table 4.9: Cross-corpus Validation Results between D-Vlog and LMVD Datasets on Multimodal Features.

Fusion	Train	Test	WAA	WAP	WAR	WAF1	UAA	UAP	UAR	UAF1
LT	D-Vlog	LMVD	0.6617 ± 0.083	0.8319 ± 0.047	0.9016 ± 0.078	0.8529 ± 0.019	0.6668 ± 0.071	0.8249 ± 0.054	0.8986 ± 0.079	0.8472 ± 0.024
	LMVD	D-Vlog	0.6367 ± 0.075	0.8131 ± 0.064	0.9040 ± 0.160	0.8188 ± 0.120	0.5816 ± 0.044	0.8427 ± 0.056	0.9166 ± 0.139	0.8456 ± 0.106
IT	D-Vlog	LMVD	0.7170 ± 0.093	0.7997 ± 0.197	0.8529 ± 0.139	0.8183 ± 0.177	0.7167 ± 0.084	0.7981 ± 0.191	0.8514 ± 0.131	0.8161 ± 0.170
	LMVD	D-Vlog	0.6856 ± 0.031	0.8222 ± 0.021	0.9009 ± 0.050	0.8561 ± 0.017	0.6462 ± 0.045	0.8498 ± 0.022	0.9154 ± 0.044	0.8781 ± 0.020
IA	D-Vlog	LMVD	**0.7454 ± 0.042**	**0.8784 ± 0.038**	0.8736 ± 0.047	**0.8715 ± 0.020**	**0.7382 ± 0.043**	**0.8756 ± 0.034**	0.8661 ± 0.058	0.8660 ± 0.028
	LMVD	D-Vlog	0.6751 ± 0.028	0.8118 ± 0.022	**0.9172 ± 0.050**	0.8562 ± 0.013	0.6265 ± 0.043	0.8414 ± 0.022	**0.9294 ± 0.042**	**0.8784 ± 0.016**

features effectively. Additionally, the features in D-Vlog appeared more robust for cross-corpus testing, possibly because of its more diverse and realistic content than LMVD. This shows that D-Vlog can significantly enhance research on depression detection in several contexts.

4.5 Summary

The experimental results presented in this chapter demonstrate that both proposed architectures achieve substantial improvements over existing methods for multimodal depression detection from social media video blogs. MDD-Net reaches an F1-Score of 77.07% on the D-Vlog dataset, surpassing prior methods by margins ranging from 1.82% to 17.37%. The ablation studies confirm that the mutual transformer fusion is responsible for this improvement, outperforming conventional baselines such as element-wise addition, multiplication, and concatenation. The weight visualizations and t-SNE plots further illustrate that the fused representations learned by MDD-Net produce distinguishable clusters for the depressed and normal classes, though some overlap remains due to the inherent noise in subjectively labeled social media data.

MMFformer pushes the performance further, achieving F1-Scores of 90.92% on D-Vlog and 90.48% on LMVD, representing improvements of 13.92% and 7.74% over the best previously reported results. Among the three proposed fusion strategies, the IT fusion performs best on D-Vlog, while the LT fusion achieves the highest scores on LMVD, suggesting that the optimal fusion level depends on the characteristics of the dataset and its feature extraction pipeline. The cross-corpus validation experiments confirm that the IA fusion generalizes most effectively across datasets, maintaining strong performance even when the training and test data come from different social media platforms, with distinct recording conditions and user demographics.

These findings carry broader implications for the field of multimodal affective computing. The consistent superiority of cross-attention-based fusion over symmetric operations supports the hypothesis that directional, modality-aware interactions are essential for capturing the complementary nature of audio and visual depression cues. This observation is consistent with recent trends in the wider multimodal AI community, where VLMs and MLLMs have shown that cross-modal attention mechanisms are more effective than late or simple concatenation-based fusion for tasks ranging from medical image analysis [60, 61] to emotion recognition [47, 48]. Comprehensive surveys on MLLMs in medicine [62, 63] further highlight the growing role of unified multimodal architectures in clinical applications, reinforcing the relevance of the fusion strategies explored in this thesis.

Chapter 5

Conclusion and Future Work

This chapter summarizes the key findings of this thesis, discusses the limitations of the current work, and outlines directions for future research.

5.1 Summary of Contributions

This thesis presented two transformer-based multimodal fusion architectures for automated depression detection from social media video blog data. The research addressed the fundamental challenge of effectively extracting and combining acoustic and visual features from noisy, uncontrolled social media content for depression classification.

In the first part of this work, MDD-Net was introduced as a multimodal depression detection system. The model uses mutual transformers to combine acoustic and visual features, thereby enhancing performance. The mutual transformer is designed to fuse the extracted features from the acoustic and visual modules for consolidating the correlation among multiple modalities. Experiments are conducted on the latest social media-based multimodal D-Vlog dataset, and the proposed network achieves comparatively better performance than the literature methods, with percentage improvements in F1-Score ranging from 1.82% to 17.37%.

In the second part, a multimodal fusion network, MMFformer, is proposed for detecting depression across multiple modalities, including video and audio. The video data is processed through a transformer network along with residual connections to extract spatial information. The audio data is exploited by a transformer encoder that helps preserve important information over time, allowing the model to capture significant temporal patterns efficiently. Moreover, the proposed system incorporates multimodal capabilities, combining features from multiple modalities

through late and intermediate fusion strategies. Experiments on two benchmark datasets reveal that the developed architecture outperforms existing methods in terms of precision of 89.24% and 90.13%, recall of 93.80% and 91.12%, and F1-Score of 90.92% and 90.48% on the D-Vlog and LMVD datasets, respectively.

Together, MDD-Net and MMFformer demonstrate that carefully designed cross-modal transformer fusion can substantially advance the SOTA in automated depression screening from social media content. The mutual transformer in MDD-Net and the three fusion strategies in MMFformer each offer distinct advantages depending on the dataset characteristics, confirming the importance of task-specific fusion mechanisms for multimodal depression detection.

5.2 Limitations

Despite the promising results achieved by both proposed architectures, several limitations should be acknowledged. First, depression datasets often include limited amounts of data and imbalanced data distributions, posing difficulties for real-world applications. Given that the D-Vlog dataset is subjectively labeled by humans, mislabeling is inevitable, increasing the risk of overfitting and limiting the network's ability to extract robust features.

Both architectures rely on pre-extracted features, specifically OpenSMILE eGeMAPS descriptors for audio and Dlib facial landmarks for video, rather than operating on raw data. While this approach reduces computational requirements, it constrains the models to a fixed set of handcrafted features and prevents them from learning task-specific low-level representations directly from the raw signals.

Additionally, the study is limited to binary classification of depression versus non-depression and does not address depression severity estimation, which would be more clinically useful. The evaluation is restricted to the D-Vlog and LMVD datasets, which, despite being the largest available benchmarks, may not fully represent the diversity of depression manifestations across different cultural and linguistic backgrounds.

Finally, neither architecture incorporates text-based analysis of speech content, which could provide additional discriminative information. The absence of textual modality means that important linguistic cues such as word choice, speech content, and sentiment patterns are not captured by the current frameworks.

5.3 Future Work

In the future, more robust and reliable multimodal depression-detection architectures should be developed and tested across multiple datasets to minimize the impact of mislabeled samples and ensure broader applicability. Moreover, we plan to evaluate the proposed MDD-Net architecture against various fusion approaches, including late, intermediate, and early fusion, to assess its efficiency in enhancing cross-modal feature integration for multimodal depression detection.

Several additional directions merit exploration in future work. First, extending the architectures to handle raw audio and video inputs through end-to-end learning could allow the models to discover task-specific features beyond the limitations of pre-extracted descriptors. Secondly, incorporating text modality via automatic speech recognition and natural language processing could provide complementary linguistic cues for depression detection, as demonstrated by recent work that integrates acoustic landmarks into LLMs [51] and uses LLM-empowered graph structures for clinical interview analysis [50].

Third, the rapid advancement of VLMs and MLLMs opens a particularly promising direction for multimodal depression screening. Recent unified architectures such as NExT-GPT [55] support any-to-any multimodal generation across text, image, video, and audio, while Chameleon [64] treats all modalities as discrete tokens within a single early-fusion transformer. These developments suggest that depression detection could benefit from large-scale foundation models capable of jointly reasoning across facial expressions, vocal patterns, and linguistic content, without requiring separate feature-extraction pipelines. The success of domain-adapted VLMs in healthcare, including Med-Gemini [61] for medical reasoning and CONtrastive learning from Captions for Histopathology (CONCH) [60] for computational pathology, indicates that similar adaptation strategies could be applied to mental health screening. MentaLLaMA [49] has already demonstrated the feasibility of instruction-tuned LLMs for interpretable mental health analysis on social media text, and extending such approaches to incorporate audiovisual inputs from video blogs represents a natural next step.

Fourth, adapting the frameworks for estimating depression severity on a continuous scale rather than a binary classification would increase their clinical relevance. Fifth, evaluating the proposed approaches on datasets from diverse cultural and linguistic backgrounds would strengthen the evidence for their generalizability. Finally, investigating the deployment of these architectures in real-time screening applications, while addressing the associated ethical and privacy considerations, represents an important step toward practical impact.

References

[1] World Health Organization. Depression. https://www.who.int/news-room/fact-sheets/detail/depression, 2023. Accessed: 2024.

[2] Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. The PHQ-9: Validity of a Brief Depression Severity Measure. *Journal of General Internal Medicine*, 16(9):606–613, 2001.

[3] Aaron T Beck, Calvin H Ward, Mock Mendelson, Jeremiah Mock, and John Erbaugh. An Inventory for Measuring Depression. *Archives of General Psychiatry*, 4(6):561–571, 1961.

[4] Wheidima Carneiro De Melo, Eric Granger, and Abdenour Hadid. Depression Detection Based on Deep Distribution Learning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4544–4548. IEEE, 2019.

[5] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting Depression via Social Media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, pages 128–137, 2013.

[6] Stevie Chancellor and Munmun De Choudhury. Methods in Predictive Techniques for Mental Health Status on Social Media: A Critical Review. *NPJ Digital Medicine*, 3(1):43, 2020.

[7] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition. In *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*, pages 3–12, 2019.

[8] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. A Review of Depression and Suicide Risk Assessment using Speech Analysis. *Speech Communication*, 71:10–49, 2015.

[9] Jeffrey M Girard, Jeffrey F Cohn, Mohammad H Mahoor, S Mohammad Mavadati, Zakia Hammal, and Dean P Rosenwald. Nonverbal Social Withdrawal in Depression: Evidence from Manual and Automatic Analyses. *Image and Vision Computing*, 32(10):641–647, 2014.

[10] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018.

[11] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, 2017.

[12] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, 2019.

[13] Graham Thornicroft, Somnath Chatterji, Sara Evans-Lacko, Michael Gruber, Nancy Sampson, Sergio Aguilar-Gaxiola, Ali Al-Hamzawi, Jordi Alonso, Laura Andrade, Guilherme Borges, et al. Undertreatment of People with Major Depressive Disorder in 21 Countries. *The British Journal of Psychiatry*, 210(2):119–124, 2017.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All You Need. *Advances in Neural Information Processing Systems*, 30, 2017.

[15] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *Advances in Neural Information Processing Systems*, 32, 2019.

[16] Jeewoo Yoon, Chaewon Kang, Seungbae Kim, and Jinyoung Han. D-Vlog: Multimodal Vlog Dataset for Depression Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12226–12234, 2022.

[17] Lang He, Kai Chen, Junnan Zhao, Yimeng Wang, Ercheng Pei, Haifeng Chen, Jiewei Jiang, Shiqing Zhang, Jie Zhang, Zhongmin Wang, et al. LMVD: A Large-Scale Multimodal Vlog Dataset for Depression Detection in the Wild. *Information Fusion*, page 103632, 2025.

[18] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the Munich Versatile and Fast Open-source Audio Feature Extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.

[19] Davis E King. Dlib-ml: A Machine Learning Toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009.

[20] Zhuoran Shen, Irwan Bello, Raviteja Vemulapalli, Xuhui Jia, and Ching-Hui Chen. Global Self-Attention Networks for Image Recognition. *arXiv preprint arXiv:2010.03019*, 2020.

[21] Yun Liu, Yu-Huan Wu, Guolei Sun, Le Zhang, Ajad Chhatkuli, and Luc Van Gool. Vision Transformers with Hierarchical Attention. *Machine Intelligence Research*, 21(4):670–683, 2024.

[22] Kateryna Chumachenko, Alexandros Iosifidis, and Moncef Gabbouj. MMA-DFER: MultiModal Adaptation of unimodal models for Dynamic Facial Expression Recognition in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4673–4682, 2024.

[23] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. *arXiv:2104.01778*, 2021.

[24] Waleed Bin Tahir, Shah Khalid, Sulaiman Almutairi, Mohammed Abohashrh, Sufyan Ali Memon, and Jawad Khan. Depression Detection in Social Media: A Comprehensive Review of Machine Learning and Deep Learning Techniques. *IEEE Access*, 13:12789–12818, 2025.

[25] Afreen Subuhi and Muniraju Naidu Vadlamudi. Advancements in AI for Depression Detection: A Survey of Machine Learning and Deep Learning Applications. In *IEEE 4th International Conference on ICT in Business Industry & Government (ICTBIG)*, pages 1–12. IEEE, 2024.

[26] Mandar Deshpande and Vignesh Rao. Depression Detection using Emotion Artificial Intelligence. In *International Conference on Intelligent Sustainable Systems (ICISS)*, pages 858–862. IEEE, 2017.

[27] Ah Young Kim, Eun Hye Jang, Seung-Hwan Lee, Kwang-Yeon Choi, Jeon Gue Park, and Hyun-Chool Shin. Automatic Depression Detection using Smartphone-Based Text-Dependent Speech Signals: Deep Convolutional Neural Network Approach. *Journal of Medical Internet Research*, 25:e34474, 2023.

[28] Kübra Erat, Elif Bilge Şahin, Furkan Doğan, Nur Merdanoğlu, Ahmet Akcakaya, and Pınar Onay Durdu. Emotion Recognition with EEG-Based Brain-Computer Interfaces: A Systematic Literature Review. *Multimedia Tools and Applications*, pages 1–48, 2024.

[29] Xiuzhuang Zhou, Kai Jin, Yuanyuan Shang, and Guodong Guo. Visually Interpretable Representation Learning for Depression Recognition from Facial Images. *IEEE Transactions on Affective Computing*, 11(3):542–552, 2018.

[30] Mazhar Hussain, Mattias O'Nils, Jan Lundgren, and Seyed Jalaleddin Mousavirad. A Comprehensive Review On Deep Learning-Based Data Fusion. *IEEE Access*, 12:180093–180124, 2024.

[31] Kyungeun Min, Jeewoo Yoon, Migyeong Kang, Daeun Lee, Eunil Park, and Jinyoung Han. Detecting Depression on Video Logs using Audiovisual Features. *Humanities and Social Sciences Communications*, 10(1):1–8, 2023.

[32] Li Zhou, Zhenyu Liu, Zixuan Shangguan, Xiaoyan Yuan, Yutong Li, and Bin Hu. TAMFN: Time-Aware Attention Multimodal Fusion Network for Depression Detection. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:669–679, 2022.

[33] Nadee Seneviratne and Carol Espy-Wilson. Multimodal Depression Classification using Articulatory Coordination Features and Hierarchical Attention Based Text Embeddings. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6252–6256. IEEE, 2022.

[34] Chengbo Yuan, Xuxu Liu, Qianhui Xu, Yongqian Li, Yong Luo, and Xin Zhou. Depression Diagnosis and Analysis via Multimodal Multi-Order Factor Fusion. In *International Conference on Artificial Neural Networks*, pages 56–70. Springer, 2024.

[35] Ying Shen, Huiyu Yang, and Lin Lin. Automatic Depression Detection: An Emotional Audio-Textual Corpus and a GRU/BiLSTM-based Model. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6247–6251. IEEE, 2022.

[36] Tianfei Ling, Deyuan Chen, and Baobin Li. MDAVIF: A Multi-Domain Acoustical-Visual Information Fusion Model for Depression Recognition from Vlog Data. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8115–8119. IEEE, 2024.

[37] Li Zhou, Zhenyu Liu, Xiaoyan Yuan, Zixuan Shangguan, Yutong Li, and Bin Hu. CAIINET: Neural Network Based on Contextual Attention and Information Interaction Mechanism for Depression Detection. *Digital Signal Processing*, 137:103986, 2023.

[38] Jiaxin Ye, Junping Zhang, and Hongming Shan. Depmamba: Progressive fusion mamba for multimodal depression detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[39] Tao Xing, Yutao Dou, Xiaolan Xie, Jiansong Zhou, Xianliang Chen, and Shaoliang Peng. EMO-Mamba: Multimodal Selective Structured State Space Model for Depression Detection. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2726–2731. IEEE, 2024.

[40] Zixuan Shangguan, Xiaxi Li, Yanjie Dong, and Xiaoyan Yuan. Automatic Depression Detection Using Attention-Based Deep Multiple Instance Learning. In *International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness*, pages 40–51. Springer, 2023.

[41] Mateusz Kowalewski, Maciej Stroinski, Kamil Kwarciak, Volodymyr Laptiev, and Daria Hemmerling. End-to-End Multimodal System for Depression Detection from Online Recordings. In *45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–4. IEEE, 2023.

[42] Lang He, Junnan Zhao, Jie Zhang, Jiewei Jiang, Senqing Qi, Zhongmin Wang, and Di Wu. LMTformer: facial depression recognition with lightweight multi-scale transformer from videos. *Applied Intelligence*, 55(2):195, 2025.

[43] Lang He, Zheng Li, Prayag Tiwari, Cui Cao, Jize Xue, Feng Zhu, and Di Wu. Depressformer: Leveraging video swin transformer and fine-grained local features for depression scale estimation. *Biomedical Signal Processing and Control*, 96:106490, 2024.

[44] Yongfeng Tao, Minqiang Yang, Huiru Li, Yushan Wu, and Bin Hu. DepMSTAT: Multimodal spatio-temporal attentional transformer for depression detection. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):2956–2966, 2024.

[45] Yongfeng Tao, Minqiang Yang, Yushan Wu, Kevin Lee, Adrienne Kline, and Bin Hu. Depressive Semantic Awareness from Vlog Facial and Vocal Streams via Spatio-Temporal Transformer. *Digital Communications and Networks*, 10(3):577–585, 2024.

[46] Minqiang Yang, Yueze Liu, Yongfeng Tao, and Bin Hu. Spike Memory Transformer: An Energy-Efficient Model in Distributed Learning Framework for Autonomous Depression Detection. *IEEE Internet of Things Journal*, 2025.

[47] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, et al. Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction Tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 110805–110853, 2024.

[48] Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, Jiangyan Yi, and Jianhua Tao. AffectGPT: A New Dataset, Model, and Benchmark for Emotion Understanding with Multimodal Large Language Models. In *Forty-second International Conference on Machine Learning*, 2025.

[49] Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. MentaLLaMA: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM Web Conference 2024*, pages 4489–4500, 2024.

[50] Zhuang Chen, Jiawen Deng, Jinfeng Zhou, Jincenzi Wu, Tieyun Qian, and Minlie Huang. Depression detection in clinical interviews with LLM-empowered structural element graph. In *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers)*, pages 8181–8194, 2024.

[51] Xiangyu Zhang, Hexin Liu, Kaishuai Xu, Qiquan Zhang, Daijiao Liu, Beena Ahmed, and Julien Epps. When LLMs Meet Acoustic Landmarks: An Efficient Approach to Integrate Speech into Large Language Models for Depression Detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 146–158, 2024.

[52] Misha Sadeghi, Robert Richer, Bernhard Egger, Lena Schindler-Gmelch, Lydia Helene Rupp, Farnaz Rahimi, Matthias Berking, and Bjoern M Eskofier. Harnessing Multimodal Approaches for Depression Detection using Large Language Models and Facial Expressions. *npj Mental Health Research*, 3(1):66, 2024.

[53] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention Augmented Convolutional Networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3286–3295, 2019.

[54] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When Does Label Smoothing Help? *Advances in Neural Information Processing Systems*, 32, 2019.

[55] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. NExT-GPT: Any-to-Any Multimodal LLM. In *Forty-first International Conference on Machine Learning*, 2024.

[56] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-IO 2: Scaling autoregressive multimodal models with vision, language, audio, and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26439–26455, 2024.

[57] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One Single Transformer to Unify Multimodal Understanding and Generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[58] Shi Yin, Cong Liang, Heyan Ding, and Shangfei Wang. A Multi-Modal Hierarchical Recurrent Neural Network for Depression Detection. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 65–71, 2019.

[59] George Dimitriadis, Joana P Neto, and Adam R Kampff. t-SNE Visualization of Large-Scale Neural Recordings. *Neural Computation*, 30(7):1750–1774, 2018.

[60] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A Visual-Language Foundation Model for Computational Pathology. *Nature medicine*, 30(3):863–874, 2024.

[61] Khaled Saab, Tao Tu, Wei-Hung Weng, et al. Capabilities of Gemini Models in Medicine. *arXiv:2404.18416*, 2024.

[62] Hanguang Xiao, Feizhong Zhou, Xingyue Liu, Tianqi Liu, Zhipeng Li, Xin Liu, and Xiaoxuan Huang. A Comprehensive Survey of Large Language Models and Multimodal Large Language Models in Medicine. *Information Fusion*, 117:102888, 2025.

[63] Rawan AlSaad, Alaa Abd-Alrazaq, Sabri Boughorbel, Arfan Ahmed, Max-Antoine Renault, Rafat Damseh, and Javaid Sheikh. Multimodal Large Language Models in Health Care: Applications, Challenges, and Future Outlook. *Journal of Medical Internet Research*, 26:e59505, 2024.

[64] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

APPENDICES

Appendix A

Architecture Configurations and Model Comparison

This appendix provides detailed reference tables summarizing the architecture configurations of both proposed frameworks, along with a comparative overview of their key design differences.

A.1 Architecture Configuration of MDD-Net

Table A.1 presents the detailed architecture configuration of MDD-Net, including the feature dimensions at each processing stage as specified in the paper.

Table A.1: Architecture Configuration of MDD-Net.

Component	**Specification**
Acoustic Feature Extraction Module (AFEM)	
Input Features	25 OpenSMILE eGeMAPS descriptors
Backbone	Global Self-Attention Network
Q, K Dimensions	$\mathbb{R}^{t \times d_a}$, $d_a = 71$
V Dimensions	$\mathbb{R}^{t \times d_o}$
Attention Type	Content + Positional
Output	$X_A^o \in \mathbb{R}^{256 \times 71}$
Visual Feature Extraction Module (VFEM)	
Input Features	68 Dlib facial landmarks (136-dim)
Backbone	HAT-Net (Hierarchical Attention)
Patch Embedding	1×1 Conv + GroupNorm + SiLU
Attention	H-MHSA with $\beta^q, \beta^k, \beta^v \in \mathbb{R}^{d_v \times d_v}$
Transformer Stages	$l_{(j-1)}$ repeats, $j \in \{2, 3, 4, 5\}$
Output	$X_V^o \in \mathbb{R}^{256 \times 139}$
Mutual Transformer (MT)	
MC_{AV}	$\mathbb{R}^{n \times d}$, audio query $\rightarrow$ visual key/value
MC_{VA}	$\mathbb{R}^{n \times d}$, visual query $\rightarrow$ audio key/value
$MC_{f_{AV}}$	$\mathbb{R}^{2n \times (d_a + d_v)}$, joint transformer
Fused Output Z	$\mathbb{R}^{4n \times 2(d_a + d_v)} = \mathbb{R}^{1024 \times 420}$
Detection Layer	
Attention	FC layer $\rightarrow$ attention scores α
Output	$p \in \mathbb{R}^2$ (Depression / Normal)
Loss Function	
Components	BCE + Focal Loss + L2 Regularization
Label Smoothing	ϵ_s (smoothing factor)
Focal Loss	Scaling ϕ, focusing γ
L2 Regularization	Strength λ

A.2 Architecture Configuration of MMFformer

Table A.2 presents the detailed architecture configuration of MMFformer.

Table A.2: Architecture Configuration of MMFformer.

Component	**Specification**
Video Feature Extraction	
Input	$\mathcal{X}_v \in \mathbb{R}^{\mathcal{T} \times C}$
Downsampling	1D Conv + Normalization + Pooling
Patch Embedding	$\mathcal{W}_{emb} \in \mathbb{R}^{C \times D}$
Special Tokens	Learnable CLS token
Positional Encoding	$\mathcal{P} \in \mathbb{R}^{1 \times (\mathcal{L}+1) \times D}$
Backbone	Pre-trained ViT with residual connections
Transformer Blocks	$\mathcal{N}$ blocks, each with MHSA + MLP
Residual	$\mathcal{X}_v^{(n+1)} = \mathcal{F}_{mlp}(\mathcal{Z}^n + \mathcal{X}_v^{(n)}) + (\mathcal{Z}^n + \mathcal{X}_v^{(n)})$
Output	$\mathcal{X}_v^{(o)} \in \mathbb{R}^{(\mathcal{L}+1) \times D}$
Audio Feature Extraction	
Input	$\mathcal{X}_a \in \mathbb{R}^S$ (raw waveform)
Projection	Conv1D + BatchNorm + Adaptive AvgPool
Patch Embedding	2D Conv with sizes (p_f, p_t), strides (s_f, s_t)
Patches	$\mathcal{M} = h \times w$
Special Tokens	CLS token + Distillation token
Positional Encoding	AST-style bilinear interpolation
Backbone	Transformer Encoder (pre-norm)
Output	$\mathcal{X}_a^{(o)} \in \mathbb{R}^{(\mathcal{M}+2) \times D}$
Fusion Module	
Late Transformer (LT)	Cross-attn after full processing
Intermediate Trans. (IT)	Cross-attn at intermediate Conv1D stage
Intermediate Attn. (IA)	Dot-product attn at intermediate stage
Cross-modal Direction	Bidirectional: $V \rightarrow A$ and $A \rightarrow V$
Post-fusion	Pooling $\rightarrow$ Concatenation $\rightarrow$ Classifier
Detection Layer	
Output	$p \in \mathbb{R}^2$ (Depression / Normal)

A.3 Comparative Summary of Model Configurations

Table A.3 provides a side-by-side comparison of the key architectural and experimental differences between MDD-Net and MMFformer to facilitate a clear understanding of how the two proposed frameworks relate to each other.

Table A.3: Comparative Summary of MDD-Net and MMFformer Configurations.

Aspect	**MDD-Net**	**MMFformer**
Sequence Length	$n = 256$	$\mathcal{L}$ (video), $\mathcal{M}$ (audio)
Audio Embed. Dim	$d_a = 71$	$\mathcal{D}$
Visual Embed. Dim	$d_v = 139$	$\mathcal{D}$
Audio Input Dim	25 (eGeMAPS)	Raw waveform
Visual Input Dim	136 (landmarks)	Landmarks / FAU
Audio Backbone	GSA-Network	AST-based Encoder
Visual Backbone	HAT-Net	ViT with Residual
Pre-trained	No	Yes (Visual MAE)
Fusion	Mutual Transformer	LT / IT / IA
Loss Function	BCE + Focal + L2	Standard BCE
Datasets	D-Vlog	D-Vlog + LMVD
Best F1 (D-Vlog)	0.7707	0.9092
Best F1 (LMVD)	–	0.9048