

BOTTLENECK CROSS-ATTENTION MAMBA NETWORK FOR MULTIMODAL DEPRESSION DETECTION

A Thesis Submitted in Partial Fulfillment of the Requirements for the
Degree of

**MASTER OF TECHNOLOGY
in
ARTIFICIAL INTELLIGENCE**

by

JYOTI PANDEY
(Roll No - 2K24/AFI/10)

Under the Supervision of

Dr. Kavinder Singh
Assistant Professor



Department of Computer Science and Engineering
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042, India

June, 2026

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

The successful completion of any task would be incomplete without acknowledging the contributions of those whose support and guidance made it possible. This project would have remained an idea without their invaluable assistance and encouragement.

Firstly, I would like to extend my heartfelt gratitude to the esteemed faculty members of our department. I am particularly thankful to my supervisor, Dr. Kavinder Singh, Department of Computer Science and Engineering, Delhi Technological University, for his unwavering encouragement and invaluable insights. His mentorship has played a vital role in shaping the idea and execution of this M.Tech Project. His support and availability for guidance at every step have been crucial to overcoming challenges along the way.

Additionally, I would like to express my sincere gratitude to everyone who directly or indirectly contributed to the completion of this project. Their assistance and encouragement have been a source of strength throughout this journey.

Lastly, I extend my deepest thanks to my parents and well-wishers for their constant support, belief in me, and encouragement, which have been my foundation in pursuing this endeavor.

Place: Delhi
Date: May 29, 2026

Jyoti Pandey
(2K24/AFI/10)

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CANDIDATE’S DECLARATION

I, JYOTI PANDEY (2K24/AFI/10), student of M.Tech in Artificial Intelligence, hereby declare that the Project titled — “Bottleneck Cross-Attention Mamba Network for Multimodal Depression Detection” which is submitted by me to the Department of Computer Science and Engineering, Delhi Technological University, New Delhi in fulfilment of the requirement for awarding of the Master of Technology degree, is not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma, Fellowship or other similar title or recognition.

Place: Delhi
Date: May 29, 2026

Jyoti Pandey
(2K24/AFI/10)

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled “Bottleneck Cross-Attention Mamba Network for Multimodal Depression Detection” which is submitted by Jyoti Pandey, Roll No - 2K24/AFI/10, Computer Science and Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the student under my supervision.

To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi
Date: May 29, 2026

Dr. Kavinder Singh
Assistant Professor

ABSTRACT

Depression detection using artificial intelligence has gained significant attention due to the limitations of traditional clinical assessment methods. This project explores the use of Mamba-based State Space Models for efficient and scalable multimodal depression detection. We implement and analyse BCM-Net, a progressive fusion architecture designed to model long-range temporal dependencies in audio-visual data while maintaining low computational complexity. By evaluating baseline models and studying their performance on large-scale benchmark datasets, this work aims to identify research opportunities for improving accuracy, efficiency, and fusion strategies. A key objective is to develop lightweight and accessible depression detection models, enabling practical deployment in real-world, resource-constrained settings.

Keywords: Depression Detection, Mamba, Selective State Space Models, Multimodal Fusion, Bottleneck Cross-Attention, Audio-Visual Analysis, BCM-Net, Mental Health Screening

CONTENTS

Acknowledgement	i
Candidate's Declaration	ii
Certificate	iii
Abstract	iv
List of Tables	vii
List of Figures	viii
List of Abbreviations	ix
List of Symbols	x
1 Introduction	1
2 Literature Survey	4
3 Methodology	9
3.1 Datasets and Data Preparation	9
3.2 Theoretical Foundations	11
3.3 Proposed BCM-Net Architecture	20
3.4 Training Protocol	25
3.5 Evaluation Strategy	26
3.6 Implementation Details	27
4 Results and Observations	28
4.1 Experimental Setup	28
4.2 Quantitative Results	28
4.3 Computational Efficiency Analysis	30
4.4 Observations and Discussion	31

5 Conclusion and Future Work	34
5.1 Conclusion	34
5.2 Future Work	35
References	35

LIST OF TABLES

3.1	Mamba vs. Transformer Self-Attention	19
3.2	BCM-Net Hyperparameter Configuration	27
4.1	Comparative Results on D-Vlog Dataset	29
4.2	Comparative Results on LMVD Dataset	29
4.3	Performance Stability Across Datasets	30
4.4	Theoretical Complexity Comparison Across Models	30
4.5	Measured Parameters and FLOPs (Available Models)	31

LIST OF FIGURES

3.1	Selective State Space Model (Mamba) [31	18
3.2	BCM-Net Architecture	20

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
AMP	Automatic Mixed Precision
AUC	Area Under the Curve
AVEC	Audio/Visual Emotion Challenge
BCA	Bottleneck Cross-Attention
BiLSTM	Bidirectional Long Short-Term Memory
CAD	Computer-Aided Diagnosis
CBA	Contextual Bi-Attention
CNN	Convolutional Neural Network
CUDA	Compute Unified Device Architecture
DTU	Delhi Technological University
EEG	Electroencephalography
eGeMAPS	Extended Geneva Minimalistic Acoustic Parameter Set
FLOPs	Floating Point Operations
GAT	Graph Attention Network
GCN	Graph Convolutional Network
GELU	Gaussian Error Linear Unit
HiPPO	High-order Polynomial Projection Operator
HBM	High Bandwidth Memory
LMVD	Large-Scale Multimodal Video Dataset
LLD	Low-Level Descriptor
MDD	Major Depressive Disorder
MHA	Multi-Head Attention
MLP	Multi-Layer Perceptron
NLP	Natural Language Processing
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
SSM	State Space Model
SVM	Support Vector Machine
WHO	World Health Organization
ZOH	Zero-Order Hold

LIST OF SYMBOLS

Symbol	Description
State Space Model Symbols	
$u(t)$	Scalar input signal at continuous time t
$x(t), h_t$	Hidden (latent) state vector at time t
$y(t)$	Scalar output at time t
A	State transition (system) matrix
$\bar{A}, \bar{A}_t$	Discretised / input-dependent state transition matrix
B	Input projection matrix
$\bar{B}, \bar{B}_t$	Discretised / input-dependent input projection matrix
C, C_t	Output (read-out) matrix
D	Direct feedthrough (skip-connection) scalar
Δ, Δ_t	Discretisation step size; time-scale parameter
K_{SSM}	SSM convolutional kernel (impulse response)
Attention Symbols	
Q	Query matrix
K	Key matrix
V	Value matrix
d_k	Dimension of query / key vectors
d_v	Dimension of value vectors
α_i	Normalised attention weight for position i
e_i	Raw attention score (logit / alignment score)
Architecture Symbols	
d	Hidden feature dimension (= 256)
d_{bn}	Bottleneck dimension for cross-attention
d_{state}	SSM state dimension (= 16)
d_{conv}	Convolution kernel size in Mamba block (= 4)
T	Sequence length (number of frames)
T_{max}	Maximum padded sequence length (= 1,000)
B	Batch size
N	SSM state size
n_{heads}	Number of attention heads (= 4)
h_a, h_v	Audio and visual temporal representations
g	Gating (selection) vector
f_{fused}	Gated multimodal fused representation
s	Utterance-level embedding after attention pooling
$\hat{y}$	Predicted class label
Training Symbols	
$\mathcal{L}$	Cross-entropy loss with label smoothing
μ, σ	Per-feature mean and standard deviation
ϵ	Small constant for numerical stability (= 10^{-8})
ϵ_{ls}	Label smoothing coefficient (= 0.1)
η	Learning rate
η_{max}, η_{min}	Peak and minimum learning rate values
λ	Weight decay coefficient (= 10^{-3})
β_1, β_2	AdamW first- and second-moment decay rates

CHAPTER 1

INTRODUCTION

Major Depressive Disorder (MDD) is one of the most common and disabling health problems of our time. According to the World Health Organization (WHO), depression is a leading cause of disability and contributes significantly to the global burden of disease, affecting hundreds of millions of people globally [1]. Although it is highly prevalent, diagnosing depression remains an exceptionally challenging task. Traditional diagnostic methods depend heavily on clinical interviews, self-report questionnaires, and subjective observations by psychiatrists. These methods are susceptible to human bias, demand active patient cooperation, and are highly resource-intensive [2, 3]. These limitations make it difficult to access mental health care quickly, especially during severe depressive episodes.

To address these issues, there has been a growing interest in digital phenotyping, which uses objective, computer-based methods to identify behavioral and linguistic features of depression [3, 4]. In this respect, Artificial Intelligence (AI) and especially Natural Language Processing (NLP) have become key frontiers in this domain. Existing studies have demonstrated that depressed individuals exhibit unique linguistic characteristics, such as increased usage of first-person singular pronouns, absolutist words, and negative emotion words [3, 4]. The ability to capture these complex, often long-range patterns in patient narratives holds significant promise for scalable and objective depression screening tools.

Automated depression detection has rapidly developed over the last decade. Traditional machine learning techniques, including Support Vector Machines (SVMs) with hand-crafted features, were primarily adopted in early methods [5]. With the advent of deep learning, recurrent neural networks (RNNs) and long short-term memory (LSTM) architectures became popular due to their ability to capture sequential data and contextual information [6]. These models dramatically improved performance but struggled with very long sequences.

The introduction of the Transformer architecture was a major breakthrough. The self-attention mechanisms of Transformers have become the standard for NLP tasks

and have achieved state-of-the-art results on depression detection benchmarks such as DAIC-WOZ [13–15]. Pretrained language models like BERT and RoBERTa showed unique abilities to capture deep semantic context [19]. Transformers, however, suffer from a fundamental limitation of quadratic time and memory complexity with respect to the sequence length ($\mathcal{O}(N^2)$) [20]. This is especially limiting when it comes to the detection of depression, which requires analyzing long-form data such as clinical interviews, therapy transcripts, or long histories of social media use. Transformers often truncate or segment inputs, breaking the long-range contextual dependencies crucial for accurate inference of mental health states [16].

Structured State Space Models (SSMs) have recently emerged as a promising alternative to overcome the scalability limitations of Transformers. Among these, the Mamba architecture represents a significant advance in sequence modeling based on the theory of continuous state spaces [20, 21]. Unlike attention-based architectures, Mamba relies on a Selective State Space mechanism that propagates information across sequences with linear time complexity ($\mathcal{O}(N)$), thus allowing for efficient long-context processing. It does so by allowing SSM parameters to depend on the input so that the model can dynamically decide what information to remember or forget at every time step. Mamba successfully combines the long-range context and fast inference of RNNs with the parallelizable training benefits of Transformers [20]. These properties make it particularly suitable for long-context mental health analysis, where subtle depression cues may be distributed across thousands of tokens.

This work presents BCM-Net (Bottleneck Cross-Attention Mamba Network), a specialized architecture for multimodal depression detection from audio-visual data. For BCM-Net, the key insight is that effective modeling of long-range temporal information and cross-modal fusion are not mutually exclusive, but can be achieved jointly via careful architectural choices. BCM-Net encodes the temporal dynamics of each modality independently with Mamba State Space Model blocks at linear complexity instead of full self-attention, which scales quadratically with the length of the sequence. This makes it suitable for processing long clinical recordings, where signals relevant to depression are often sparse and temporally distributed.

The primary difference of BCM-Net from existing Mamba-based methods lies in the cross-modal fusion strategy. After the temporal encoding for each modality, the model combines the audio and visual streams via a bidirectional bottleneck cross-attention mechanism, where each modality attends to the other in a compressed intermediate space. This reduces the cross-modal attention cost without sacrificing the model’s ability to capture meaningful inter-modal correlations. This is further enhanced by a subsequent gated fusion layer that learns the importance of each modality

at each point in time, considering that the relative importance of audio and visual signals varies across individuals and recording conditions. Further, an attention pooling layer reduces the temporal dimension into a single representation by attending to the frames most diagnostic for the classification decision.

The main goal of this work is to propose a new method BCM-Net to improve automatic binary depression classification results on two real-world benchmark datasets: D-Vlog and LMVD. Specifically, the study aims to achieve:

- (i) **Multimodal Temporal Modeling:** Assess whether Mamba’s linear-time encoding, performed independently per modality, captures long-range behavioral patterns associated with depression better than transformer-based baselines.
- (ii) **Efficient Cross-Modal Fusion:** Assess whether bottleneck cross-attention offers a computationally feasible alternative to full cross-attention while preserving sufficient inter-modal expressiveness for clinical classification.
- (iii) **Performance Benchmarking:** Evaluate macro-F1, accuracy, and per-class precision and recall, with a special focus on minimizing false negatives: a key consideration in any screening context where missed detections have clinical consequences.

This work represents a meaningful step toward depression detection systems that are both architecturally principled and practically deployable without the massive computational overhead associated with state-of-the-art multimodal models.

CHAPTER 2

LITERATURE SURVEY

Automatic depression detection has been a major focus of computational research over the last decade as mental health disorders become more prevalent and standard clinical screening methods show clear limitations. Early computational studies in this domain were heavily influenced by the benchmark datasets and evaluation protocols provided by the Audio/Visual Emotion Challenges (AVEC). The AVEC 2013 [1] and AVEC 2014 challenges [2] formalized depression diagnosis as a multimodal learning problem using semi-structured clinical interviews. This enabled detailed analysis of vocal acoustics, facial expressions, and somatic markers. Early models in these challenges primarily relied on supervised machine learning classifiers, such as Support Vector Machines (SVMs), operating on hand-crafted visual descriptors and vocal features. These initial frameworks demonstrated that objective behavioral markers could reflect underlying depressive states. However, they were limited by their reliance on static, aggregated statistics that discarded the fine-grained temporal patterns of speech and facial micro-expressions. Furthermore, because these early datasets were relatively small, shallow classifiers were highly prone to overfitting, prompting the need for more robust feature extraction protocols and deep learning paradigms that could automatically learn hierarchical representations from raw signals rather than relying on manual feature engineering.

Meanwhile, researchers began to explore social media as a passive, highly scalable data source for mental health monitoring. A pioneering study by De Choudhury et al. [3] demonstrated that linguistic patterns, emotional expression, and posting habits on Twitter could predict clinical depression months before formal diagnosis. Subsequent research expanded on this paradigm by leveraging large-scale web data while raising critical discussion regarding dataset bias, model interpretability, and ecological validity [4]. This approach, commonly referred to as digital phenotyping, leverages the natural language produced by individuals in their everyday lives to detect subtle shifts in mood and cognitive patterns. By monitoring these public streams, researchers identified key linguistic markers such as elevated use of first-person singular pronouns, a higher prevalence of negative emotion words, and a reduction in social engagement

vocabulary. Additionally, researchers began integrating temporal metadata, observing that irregular posting patterns and late-night activity often correlated with sleep disturbances typical of depressive episodes. These behavioral insights underscored the value of social media as an ecological sensor, although they also highlighted the pressing need for privacy-preserving machine learning frameworks to ensure user data is protected.

While early text models relied on shallow representations such as bag-of-words or simple tf-idf vectors, deep learning architectures introduced context-aware modeling. The CBA (Contextual Bi-Attention) framework [5] integrated Bidirectional LSTMs (BiLSTMs) with hierarchical attention mechanisms to extract long-term contextual relationships and emotional shifts from unstructured social media narratives. Recently, graph-based text models have emerged as a powerful alternative. The CDME-GAT [6] framework constructs semantic graphs from social media posts by combining contextual embeddings from large pretrained language models (BERT, RoBERTa, and DeBERTa) with Graph Attention Networks, yielding substantial improvements in both classification performance and operational efficiency. By modeling text documents or individual posts as nodes within a relational graph, these systems can capture complex semantic associations that transcend strict chronological boundaries. The incorporation of graph attention layers allows the network to dynamically assign weights to relationships between disparate posts, prioritizing high-risk linguistic connections while suppressing unrelated noise. This structural flexibility ensures that the model can trace thematic threads across a user’s digital footprint, capturing the gradual accumulation of depressive symptoms over extended periods. Consequently, this shift from flat sequential architectures to graph-based structures has enabled a much deeper representation of the cognitive and narrative coherence of patients.

In parallel, multimodal depression detection has consistently demonstrated superior performance compared to unimodal approaches. Deep multimodal methods originally fused audio and video streams using combinations of 2D/3D CNNs and RNNs. Temporal causal convolutional networks were subsequently introduced to capture structured sequential behavior [7]. The clinical utility of combining facial expressions with acoustic descriptors has been successfully demonstrated in psychiatric settings for relapse prediction [8] and diagnostic screening. By integrating multiple behavioral channels, these multimodal systems can offset the limitations of any single modality, such as an individual hiding facial expressions while still exhibiting flat vocal affect. However, first-generation multimodal models typically combined these channels using simple concatenation at the input level or late-stage decision voting. These naive fusion approaches often failed to model the intricate cross-modal alignments and tem-

poral synchronization between speech pauses and facial micro-expressions. As a result, researchers recognized that simple feature stacking could not capture the subtle, joint emotional expressions that characterize major depressive disorder, motivating the development of more advanced, interactive fusion paradigms.

However, these models face challenges in multimodal fusion and computational scalability. To address representation bias, Fu et al. [9] developed a graph-based multimodal model guided by Facial Action Units (FAUs) to isolate identity-independent affective cues. Similarly, Yang and Zheng proposed MSGAF (Multi-Feature and Graph Attention Fusion) [10], which utilizes bidirectional fusion graphs to represent intra-modal temporal dynamics and inter-modal dependencies. Related techniques have explored complex graph structures further, including Graph Convolutional Networks (GCNs) over speech-EEG signals [11] and hypergraph-based multimodal representations [12]. By utilizing graph and hypergraph structures, these frameworks can model multi-way relationships among vocal pitches, facial movements, and neurological waveforms. For instance, a hypergraph can group synchronous occurrences of sighing, downcast gaze, and specific EEG frequency drops into a single hyperedge, preserving multi-modal context that binary graphs discard. Nevertheless, these graph-based networks introduce substantial computational complexity during the adjacency matrix updates and message-passing phases. This overhead, combined with the extensive preprocessing required for tracking facial landmarks and building graphs, limits their suitability for real-time clinical screening or deployment on resource-constrained devices, pointing to a need for more streamlined sequence architectures.

A further major milestone was the introduction of Transformer-based architectures. MMFormer [13] proposed a multimodal fusion transformer designed to model mutual interactions between audio-visual streams at the utterance level, capturing the complex spatio-temporal alignment of human communicative behavior. This sentence-level cross-modal alignment achieved state-of-the-art results on standard benchmarks such as DAIC-WOZ, a trend reinforced by other recent architectures like MDD-Net [14] and SMFL (Sentence-level Multimodal Fusion Learning) [15]. The primary strength of these Transformer-based designs lies in their self-attention and cross-attention mechanisms, which allow the network to evaluate dependencies across different time steps and modalities without relying on recurrent step-by-step calculations. For example, a visual token representing a sudden downcast gaze can be directly aligned with a vocal token representing a whispering pause, regardless of their position in the sequence. By organizing the clinical dialogue into utterance-level segments, these models preserve the conversational flow and capture semantic nuances that traditional segment-agnostic models overlook. This capacity to capture fine-grained, bidirectional interactions be-

tween vocal characteristics and facial dynamics has set a new standard for representational accuracy in diagnostic models.

Despite their impressive representational capacity, Transformer-based models suffer from a fundamental limitation: their self-attention mechanism incurs quadratic time and memory complexity relative to the sequence length. This constraint heavily limits their ability to process long-form clinical recordings or extensive histories of social media activity. To address this limitation, researchers have recently turned to Mamba-based State Space Models (SSMs). Because standard self-attention requires computing a dense dot-product matrix between all pairs of time steps, the computational cost scales quadratically as the session length increases. In clinical depression screening, where interviews can exceed thirty minutes and require high-frequency sampling to capture micro-expressions, this quadratic scaling leads to severe GPU memory exhaustion. The Mamba architecture resolves this bottleneck by utilizing a selective scan mechanism that models long-range sequences with linear time complexity. This approach enables the network to maintain a recurrent, constant-time inference mode while preserving the capacity to capture long-term contextual dependencies, making it a highly attractive alternative for processing extended behavioral streams.

Ye et al. [16] were the first to apply state space models to multimodal depression detection through DepMamba, which implements a progressive state space fusion mechanism combined with convolutional feature extraction to model long-range temporal dependencies efficiently. MDDMamba [17] subsequently introduced a memory-efficient cross-modal attention mechanism built on Mamba blocks to reduce fusion costs. This approach was further extended by STE-Mamba [18], which combines emotional trend analysis with spatio-temporal selective state space models to achieve state-of-the-art accuracy. These state-space frameworks represent a paradigm shift in how long-form clinical interviews are analyzed, demonstrating that sequence modeling does not have to come at the expense of computational tractability. By using hardware-aware parallel scans, they allow models to scale efficiently to thousands of frames without sacrificing accuracy. For instance, STE-Mamba models spatial facial dynamics and acoustic frequency trends in parallel selective state space blocks before fusing them, thereby capturing both micro-expressions and long-term emotional variations. These models have demonstrated substantial improvements in speed and memory efficiency, paving the way for automated screening tools that can run on standard clinical computers without high-end GPU requirements.

Recent systematic surveys and benchmarking reviews [19] underscore the growing importance of combining large-scale hybrid architectures with robust sequence models, validating the clinical utility and scalability of these systems in therapeutic

settings. The literature reveals a clear trajectory from hand-crafted statistical representations toward complex, graph-guided, Transformer-based, and State Space sequence models, driven by the dual requirements of clinical accuracy and computational scalability. Modern research is increasingly exploring the integration of clinical text transcripts processed by pre-trained language models alongside vocal and visual cues to form more comprehensive diagnostic profiles. This evolution towards hybrid systems highlights the potential of automated tools to serve as objective, passive screening aids that support clinicians during standard diagnostic sessions. To transition these models successfully from benchmarks to real-world clinics, researchers are now focusing on cross-cohort validation to ensure generalizability across diverse demographic groups. Ultimately, finding a balance between rich cross-modal alignment and linear-time computational complexity remains the primary objective, with hybrid state-space networks currently representing the state of the art in resolving this tension.

CHAPTER 3

METHODOLOGY

This chapter presents the complete methodology underlying the proposed BCM-Net (Bottleneck Cross-Attention Mamba Network) for automated multimodal depression detection. Section 3.1 describes the datasets and data preparation procedures. Section 3.2 provides a detailed account of the three core theoretical building blocks: the attention mechanism, cross-attention, and the Mamba selective state space model that underpin BCM-Net’s design. Section 3.3 presents the full six-stage BCM-Net architecture with supporting mathematical formulations. Section 3.4 covers the training protocol; Section 3.5 defines the evaluation metrics; and Section 3.6 summarises the implementation environment.

3.1 Datasets and Data Preparation

Two publicly available benchmarks are used to train and evaluate BCM-Net: the D-Vlog dataset and the Large-scale Multimodal Video Dataset (LMVD). Both provide synchronised audio-visual recordings annotated for binary depression classification (depressed / non-depressed).

3.1.1 D-Vlog Dataset

D-Vlog [22] consists of 961 YouTube vlogs collected between January 2020 and January 2021, comprising approximately 160 hours of naturalistic in-the-wild video. The corpus contains 555 depression-positive and 406 depression-negative samples recorded by 816 unique individuals. Labels were assigned according to strict linguistic criteria: a video is marked as depressed only if the vlogger explicitly discusses a current depressive episode, suicidal ideation, or ongoing pharmacotherapy; videos discussing past depression or ordinary sadness are excluded or marked non-depressed. The dataset is partitioned into training, validation, and test splits (approximately 70 / 10 / 20) using the fold labels provided in the accompanying labels.csv file. The naturalistic recording conditions varying lighting, camera angle, background noise, and speaking style make

D-Vlog a challenging and ecologically valid benchmark. BCM-Net follows the evaluation protocol adopted by DepMamba [16] and MDDMamba [17] on this dataset to ensure comparability.

3.1.2 LMVD Dataset

The Large-scale Multimodal Video Dataset (LMVD) [23] contains 1,823 videos sourced from multiple social media platforms, of which 908 are depression-positive and 915 are depression-negative, a near-balanced class distribution. The multi-platform origin introduces greater domain diversity than D-Vlog in terms of recording style and speaker demographics. LMVD is used to assess the generalisability of BCM-Net beyond a single source domain. The evaluation protocol follows that used by DepMSTAT [24] on this dataset.

3.1.3 Feature Extraction

Acoustic Features: eGeMAPS

Acoustic features are extracted using the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [25] via the openSMILE toolkit. eGeMAPS is a curated 25-dimensional acoustic descriptor designed for paralinguistic and affective computing research. It encompasses four families of parameters: (i) frequency-related features (fundamental frequency F0, jitter, shimmer), (ii) energy and amplitude features (loudness, harmonic-to-noise ratio), (iii) spectral features (spectral flux, alpha ratio, Hammarberg index), and (iv) cepstral features (MFCC coefficients 1-4).

Features are extracted frame-by-frame, yielding an acoustic tensor of shape $(B, T, 25)$, where B is the batch size and T is the number of frames. eGeMAPS has been adopted as the acoustic representation in DepMamba [16], STE-Mamba [18], and EMO-Mamba [26], providing a standardised baseline for comparison.

Visual Features: Facial Landmarks

Visual features are represented as 2D facial landmark coordinates extracted per frame using a face alignment library (e.g., OpenFace or MediaPipe). Each frame yields 68 landmark points, each with an (x, y) coordinate, producing a 136-dimensional vector per frame. The resulting visual tensor has shape $(B, T, 136)$.

The 68 landmarks span the full facial region, including brow landmarks (1-17), eye landmarks (18-48), nose landmarks (28-36), and mouth landmarks (49-68). Prior

work has shown that depression is associated with reduced eye blink rate, flattened brow affect, and diminished mouth movement [24], making landmark-based features clinically interpretable.

3.1.4 Data Preprocessing and Augmentation

All features are normalised using Z-score standardisation computed over the training set using Welford’s online algorithm to avoid loading the full dataset into memory:

$$z = \frac{x - \mu}{\sigma + \varepsilon} \quad (3.1)$$

where μ and σ are the per-feature global mean and standard deviation computed from the training split only, and $\varepsilon = 10^{-8}$ prevents division by zero. Validation and test features are normalised using training statistics to prevent data leakage.

Sequences are padded or truncated to a fixed length of $T_{max} = 1,000$ frames. A binary padding mask $M \in \{0, 1\}^{B \times T}$ is constructed for each batch, where $M_t = 1$ denotes a real frame and $M_t = 0$ denotes a padded position.

Two stochastic augmentation strategies are applied to the training split:

- (i) **Gaussian noise injection:** Small-amplitude noise is added independently to each feature dimension, simulating recording variability:

$$x_{aug} = x + \delta, \quad \delta \sim \mathcal{N}(0, 0.05^2) \quad (3.2)$$

- (ii) **Time masking:** A contiguous span of 10% of T frames is randomly selected and set to zero, following the SpecAugment paradigm [27]. This encourages the model to rely on distributed temporal cues rather than a fixed diagnostic window.

Neither augmentation is applied at validation or test time.

3.2 Theoretical Foundations

BCM-Net is built on three fundamental building blocks: the attention mechanism, cross-attention, and Mamba selective state space models. This section provides a self-contained theoretical treatment of each, which motivates the architectural decisions described in Section 3.3.

3.2.1 Attention Mechanism

The attention mechanism [30] is a function that allows a neural network to selectively focus on relevant parts of an input sequence when producing each element of an output sequence. Rather than summarising the entire input into a fixed-size context vector, attention computes a weighted combination of all input positions, where the weights reflect the relevance of each input position to the current output position.

Scaled Dot-Product Attention

The fundamental operation of modern attention is scaled dot-product attention. Given a query vector $q \in \mathbb{R}^{d_k}$, a set of key vectors $K \in \mathbb{R}^{N \times d_k}$, and a set of value vectors $V \in \mathbb{R}^{N \times d_v}$, the attention output for a single query is:

$$e_i = \frac{q^T k_i}{\sqrt{d_k}}, \quad \text{for } i = 1, \dots, N \quad (3.3)$$

$$\alpha_i = \frac{\exp(e_i)}{\sum_{j=1}^N \exp(e_j)} \quad (3.4)$$

$$\text{Attn}(q, K, V) = \sum_{i=1}^N \alpha_i v_i \quad (3.5)$$

where e_i is the raw attention score (also called the logit or alignment score) between the query and the i -th key, α_i is the normalised attention weight obtained by applying softmax over all N scores, and the output is the weighted sum of value vectors. In matrix form, given $Q \in \mathbb{R}^{L \times d_k}$ (L queries), $K \in \mathbb{R}^{N \times d_k}$, and $V \in \mathbb{R}^{N \times d_v}$:

$$\text{Attn}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (3.6)$$

The scaling factor $1/\sqrt{d_k}$ prevents the dot products from growing large in magnitude as d_k increases, which would push the softmax into regions of very small gradient. This scaling was introduced by Vaswani et al. [30] as a stabilising heuristic.

The attention weights α_i are interpretable: a weight close to 1.0 on position i means the output is almost entirely determined by the value at position i , while a uniform distribution indicates equal reliance on all positions.

Multi-Head Attention

A single attention head can only capture one type of positional relationship at a time. Multi-head attention (MHA) [30] projects queries, keys, and values into h parallel subspaces using separate learned projections, computes attention independently in each subspace, and concatenates the results:

$$\text{head}_i = \text{Attn}(QW_Q^i, KW_K^i, VW_V^i) \quad (3.7)$$

$$\text{MHA}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W_O \quad (3.8)$$

where $W_Q^i \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_K^i \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_V^i \in \mathbb{R}^{d_{\text{model}} \times d_v}$ are per-head projection matrices, and $W_O \in \mathbb{R}^{h \cdot d_v \times d_{\text{model}}}$ is the output projection. Typically $d_k = d_v = d_{\text{model}}/h$, so the total computation cost is the same as single-head attention.

Each head learns to attend to different types of relationships simultaneously. For example, one head may focus on local temporal patterns while another captures long-range dependencies. This parallel multi-perspective attention is a key reason for the expressiveness of Transformer-based models.

Self-Attention and Temporal Modelling

When the query, key, and value matrices all come from the same sequence, i.e., $Q = K = V = X$. The operation is called self-attention. Every position in the sequence attends to every other position, producing a context-aware representation that captures global dependencies regardless of the distance between positions.

Self-attention computes an $N \times N$ attention matrix (where N is the sequence length), giving it $\mathcal{O}(N^2)$ time and memory complexity. For short sequences this is acceptable, but for long clinical vlog recordings where N can reach hundreds to thousands of frames, the quadratic cost becomes a significant bottleneck. This limitation is one of the primary motivations for adopting Mamba (Section 3.2.4) in BCM-Net.

In the context of depression detection, self-attention has been used in models such as MMFformer [13] and DepMSTAT [24], where it enables each frame representation to be updated based on all other frames in the recording. However, these models incur high memory and compute costs for long sequences.

3.2.2 Cross-Attention

Cross-attention is a generalisation of the attention mechanism in which the query sequence and the key-value sequence come from two different sources. It enables one

modality or representation to “look into” another and retrieve relevant information from it, making it the natural mechanism for multimodal fusion.

Distinction from Self-Attention

In self-attention, Q, K, V are all derived from the same input X . In cross-attention, the queries come from one source sequence X and the keys and values come from a different source sequence Y :

$$Q = XW_Q, \quad K = YW_K, \quad V = YW_V \quad (3.9)$$

$$\text{CA}(X, Y) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (3.10)$$

The output has the same length as X (one output per query), but each output vector is informed by the content of Y through the attention weights. Intuitively, each position in X asks a “question” (query), and the attention mechanism finds the most relevant “answers” (keys) in Y , then retrieves the associated “information” (values) from Y .

Cross-attention is the core mechanism in the encoder-decoder architecture of the original Transformer [30], where the decoder queries the encoder output to condition text generation on the source sequence.

Multi-Head Cross-Attention

Multi-head cross-attention follows the same formulation as multi-head self-attention (Eq. 3.8), except Q and K, V come from different sources:

$$\text{head}_i^{X \rightarrow Y} = \text{Attn}(XW_Q^i, YW_K^i, YW_V^i) \quad (3.11)$$

$$\text{MHCA}(X, Y) = \text{Concat}(\text{head}_1^{X \rightarrow Y}, \dots, \text{head}_h^{X \rightarrow Y}) W_O \quad (3.12)$$

The multi-head formulation allows the model to capture multiple types of cross-modal relationships simultaneously: for example, in the context of depression detection, one head might correlate vocal energy (audio) with mouth movement (visual), while another correlates speech rate with brow position.

Role in Multimodal Fusion

In a multimodal setting with modalities X (e.g., audio) and Y (e.g., visual), cross-attention can be applied bidirectionally:

- **Audio attends to visual:** Each audio frame gathers context from all visual frames $CA(h_a, h_v)$.
- **Visual attends to audio:** Each visual frame gathers context from all audio frames $CA(h_v, h_a)$.

This bidirectional cross-attention ensures that information flows symmetrically between modalities, producing enriched representations for both streams that encode inter-modal context. This is the strategy adopted in BCM-Net’s Stage 3 (Section 3.3.4), and is contrasted with late fusion (where modalities are combined only at the output level) and concatenation-based early fusion (which lacks the selective alignment that attention provides).

3.2.3 State Space Models

State space models (SSMs) are a class of sequence models rooted in control theory and signal processing. They represent a sequence as a dynamical system: a hidden state evolves over time according to a differential equation, and observations are linear readouts of that state. SSMs offer a compelling alternative to self-attention because they process sequences at $\mathcal{O}(T)$ time and memory complexity rather than $\mathcal{O}(T^2)$.

Continuous-Time Linear SSMs

The continuous-time linear SSM is defined by the following differential and output equations:

$$\frac{dx(t)}{dt} = Ax(t) + Bu(t) \quad (3.13)$$

$$y(t) = Cx(t) + Du(t) \quad (3.14)$$

where:

- $u(t) \in \mathbb{R}$ is the scalar input signal at time t
- $x(t) \in \mathbb{R}^N$ is the hidden state vector of dimension N (state size)
- $y(t) \in \mathbb{R}$ is the scalar output
- $A \in \mathbb{R}^{N \times N}$ is the state transition (or system) matrix
- $B \in \mathbb{R}^{N \times 1}$ is the input projection matrix
- $C \in \mathbb{R}^{1 \times N}$ is the output projection matrix

- $D \in \mathbb{R}$ is a scalar direct feedthrough (skip connection) term

The matrix A governs how information stored in the hidden state decays or propagates over time. For stable systems, the eigenvalues of A must have negative real parts. The matrices B and C determine how the input writes into the state and how the state is read out as the output, respectively.

For discrete-time sequence modelling (as required for processing frame-level features), the continuous SSM is discretised using the zero-order hold (ZOH) rule with a step size Δ (the sampling interval):

$$\bar{A} = \exp(\Delta A) \quad (3.15)$$

$$\bar{B} = (\Delta A)^{-1}(\exp(\Delta A) - I) \cdot \Delta B \quad (3.16)$$

This yields the recurrence:

$$h_t = \bar{A}h_{t-1} + \bar{B}x_t \quad (3.17)$$

$$y_t = Ch_t + Dx_t \quad (3.18)$$

where $h_t \in \mathbb{R}^N$ is the discrete hidden state at timestep t , $x_t \in \mathbb{R}$ is the input token, and $y_t \in \mathbb{R}$ is the output token. This recurrence can be unrolled as a convolutional operation over the full sequence, enabling efficient parallel training:

$$y = K_{SSM} * u \quad (3.19)$$

where $K_{SSM} = (CB, C\bar{A}B, C\bar{A}^2B, \dots)$ is the SSM convolutional kernel (also called the impulse response), $*$ denotes convolution, and u is the full input sequence. The ability to compute SSMs both as a recurrence (linear time, constant memory, efficient inference) and as a convolution (parallelisable, efficient training) is the central computational advantage of SSMs over RNNs.

Structured State Spaces (S4)

A key challenge for classical SSMs is initialising and learning the state transition matrix A in a way that captures long-range dependencies. Gu et al. [20] addressed this through the Structured State Space (S4) model, which initialises A using the HiPPO (High-order Polynomial Projection Operators) framework. The HiPPO matrix is designed so that the hidden state $x(t)$ optimally memorises the history of $u(t)$ up to time

t in a polynomial basis:

$$A_{nk} = \begin{cases} -\sqrt{(2n+1)(2k+1)} & \text{if } n > k \\ -(n+1) & \text{if } n = k \\ 0 & \text{if } n < k \end{cases} \quad (3.20)$$

This structured initialisation gives the SSM an inductive bias towards capturing information across multiple timescales simultaneously, from local transitions to long-range temporal trends. S4 demonstrated that SSMs with this structure could match or exceed Transformer performance on long-range sequence benchmarks while maintaining $\mathcal{O}(T)$ complexity. However, S4 uses fixed (input-independent) matrices B , C , and Δ , limiting its ability to selectively filter the input based on content — a limitation resolved by Mamba.

3.2.4 Mamba: Selective State Space Models

Mamba [20] extends the S4 framework with a content-aware (selective) parameterisation that allows the model to dynamically decide which information to retain in the hidden state and which to discard. This selectivity is the key innovation that makes Mamba outperform classical SSMs on tasks requiring content-based reasoning over long sequences, such as language modelling as demonstrated in DepMamba [16] and EMO-Mamba [26] depression detection from long clinical recordings.

Selectivity Mechanism

In S4, the matrices B , C , and the step size Δ are fixed parameters learned during training and shared in all time steps. Mamba makes these parameters input-dependent; they are computed as learned functions of the current input token x_t :

$$B_t = x_t W_B, \quad C_t = x_t W_C \quad (3.21)$$

$$\Delta_t = \text{softplus}(x_t W_\Delta + b_\Delta) \quad (3.22)$$

where W_B, W_C, W_Δ are projection matrices and b_Δ is a bias vector. The discretized matrices then become timestep-specific:

$$\bar{A}_t = \exp(\Delta_t A) \quad (3.23)$$

$$\bar{B}_t = (\Delta_t A)^{-1}(\exp(\Delta_t A) - I) \cdot \Delta_t B_t \quad (3.24)$$

And the recurrence becomes:

$$h_t = \bar{A}_t h_{t-1} + \bar{B}_t x_t \quad (3.25)$$

$$y_t = C_t h_t + D x_t \quad (3.26)$$

This selective recurrence has a compelling interpretation: the step size Δ_t controls how much time passes between consecutive inputs. A large Δ_t makes $\bar{A}_t$ approach zero (the previous state is discarded and the new input dominates), while a small Δ_t preserves the previous state strongly (the model “ignores” the current input). By making Δ_t input-dependent, Mamba can learn to focus on informative timesteps (e.g., speech segments containing depressive content) and compress uninformative ones (e.g., silence or background noise).

Similarly, input-dependent B_t controls how the current input x_t is written into the state, and input-dependent C_t controls how the state is read out. Together, these three selective gates give Mamba the ability to perform content-based filtering, a capability absent in S4 but essential for processing heterogeneous clinical recordings where relevant cues are sparse and context-dependent.

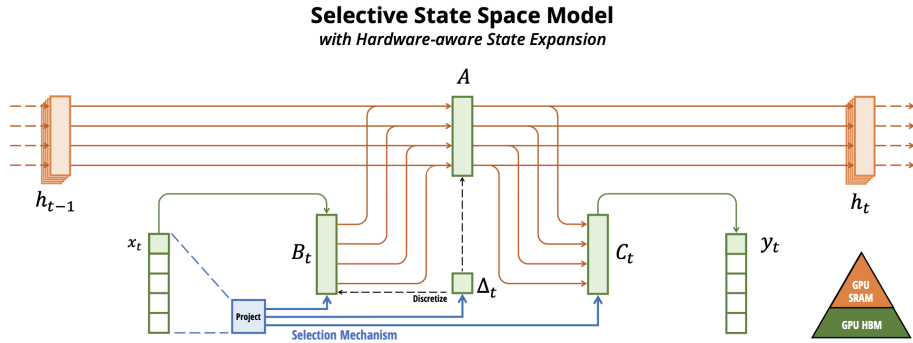


Figure 3.1.]

Selective State Space Model (Mamba) [31]: the input-dependent parameterisation of B_t , C_t , and Δ_t enables content-aware filtering of the hidden state at each timestep, allowing the model to selectively retain or discard information based on the current input token.

Hardware-Aware Parallel Scan

The selective nature of Mamba (input-dependent parameters) prevents it from being computed as a global convolution like S4. Instead, Mamba computes the recurrence using a parallel prefix scan (also known as an associative scan), which computes all elements $h_1, h_2, \dots, h_T$ in $\mathcal{O}(\log T)$ parallel steps [20].

The parallel scan is implemented in a hardware-aware manner using CUDA kernel fusion and online normalisation to avoid materialising the full $N \times T$ state matrix in high-bandwidth memory (HBM). This keeps the memory footprint at $\mathcal{O}(T \cdot N)$ rather than the $\mathcal{O}(T^2)$ of self-attention, while maintaining the expressive content-based selection of the selective scan.

Additionally, Mamba incorporates a local convolutional layer (kernel size $d_{conv} = 4$) applied to the input before the selective scan. This convolutional pre-processing captures local context within a small neighbourhood, complementing the global long-range modelling of the recurrent SSM.

Comparison with Transformers

Table 3.0 summarises the key differences between Mamba and the standard Transformer self-attention mechanism.

Table 3.1. Mamba vs. Transformer Self-Attention

Property	Transformer Attn	Mamba SSM
Time Complexity	$\mathcal{O}(T^2 \cdot d)$	$\mathcal{O}(T \cdot d \cdot N)$
Memory Complexity	$\mathcal{O}(T^2)$	$\mathcal{O}(T \cdot N)$
Parallelism (train)	Full parallel	Parallel scan
Inference mode	KV-cache $\mathcal{O}(T)$	Recurrent $\mathcal{O}(1)$
Content-awareness	Full (QK^T attention)	Selective B, C, Δ
Long-range modelling	Yes ($\mathcal{O}(T^2)$ cost)	Yes ($\mathcal{O}(T)$ cost)
Fixed context window	No	No

For long vlog sequences ($T \sim$ hundreds of frames), the quadratic memory cost of self-attention can exhaust GPU memory, while Mamba scales linearly. Empirically, Mamba has been shown to achieve $15\times$ faster inference and 92.3% lower GPU memory usage than comparable Transformer architectures at long sequence lengths [16], making it highly suitable for processing the extended audio-visual recordings in D-Vlog and LMVD.

3.3 Proposed BCM-Net Architecture

3.3.1 Architectural Overview

BCM-Net is an end-to-end binary classifier that accepts paired acoustic and visual feature streams and predicts a depression label for each recording. The architecture integrates the three theoretical components described in Section 3.2: (i) Mamba SSM blocks for efficient within-modality temporal encoding, (ii) bidirectional bottleneck cross-attention for inter-modal fusion, and (iii) attention pooling for temporally selective aggregation. A gated fusion layer provides adaptive modality weighting before classification.

The design draws on DepMamba [16] for Mamba-based multimodal depression detection, MMFformer [13] for cross-modal fusion, and MDDMamba [17] for the bottleneck cross-attention design.

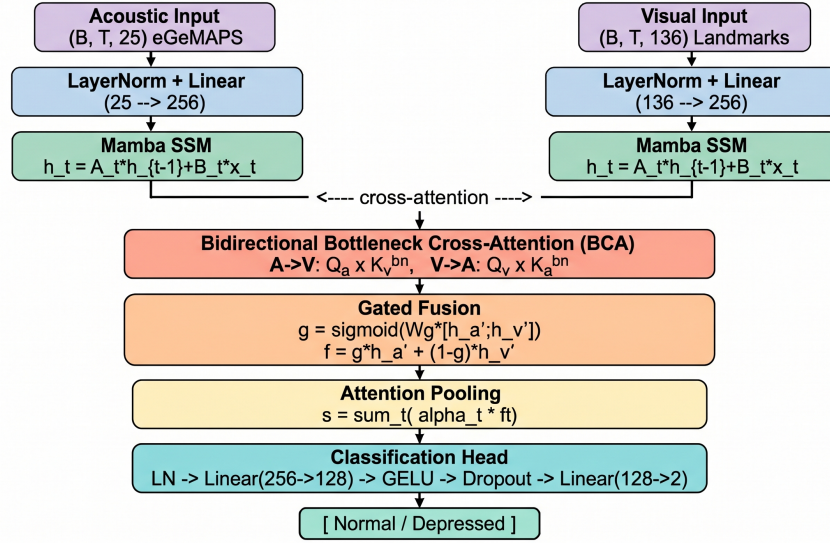


Figure 3.2. BCM-Net Architecture: the six-stage end-to-end pipeline comprising (1) input normalisation and linear projection, (2) Mamba-based within-modality temporal encoding, (3) bidirectional bottleneck cross-attention fusion, (4) gated modality weighting, (5) attention pooling, and (6) the two-layer MLP classification head.

3.3.2 Stage 1: Input Normalization and Linear Projection

Each modality is independently normalised and projected into a shared hidden space of dimension $d = 256$. Layer normalisation (LayerNorm) is applied to stabilise the

input distribution, followed by a modality-specific linear transformation:

$$h_a^{(0)} = W_a \text{LayerNorm}(x_a) + b_a \quad (3.27)$$

$$h_v^{(0)} = W_v \text{LayerNorm}(x_v) + b_v \quad (3.28)$$

where $W_a \in \mathbb{R}^{256 \times 25}$ and $W_v \in \mathbb{R}^{256 \times 136}$ are learnable weight matrices, and $b_a, b_v \in \mathbb{R}^{256}$ are bias vectors. After projection, both modalities reside in the same $\mathbb{R}^{256}$ feature space, enabling subsequent cross-modal operations. The hidden dimension $d = 256$ balances representational capacity against parameter count, consistent with comparable Mamba-based depression models [16, 17].

3.3.3 Stage 2: Mamba-Based Temporal Encoding

Independent Mamba SSM blocks (described in Section 3.2.4) are applied to $h_a^{(0)}$ and $h_v^{(0)}$ to encode within-modality temporal dynamics at $\mathcal{O}(T)$ complexity. Each modality is encoded separately at this stage, preserving modality-specific temporal structure before cross-modal fusion.

SSM Configuration

The Mamba block operates with $d_{state} = 16$, $d_{conv} = 4$, and an expansion factor of 2, following the default Mamba configuration [20]. The expansion factor projects the $d = 256$ input into an inner dimension of 512 before the selective scan, providing additional representational capacity within each block. A dropout rate of 0.3 is applied within each Mamba block for regularisation. The selective recurrence for each modality encoder is (from Eq. 3.26):

$$h_t = \bar{A}_t h_{t-1} + \bar{B}_t x_t \quad (3.29)$$

$$y_t = C_t h_t + D x_t \quad (3.30)$$

where $\bar{A}_t$, $\bar{B}_t$, and C_t are input-dependent and computed via learned linear projections of the current input token x_t .

Complexity Analysis

The Mamba selective scan achieves $\mathcal{O}(T)$ time and memory complexity, compared to $\mathcal{O}(T^2)$ for self-attention in transformer-based models such as DepMSTAT [24] and MMFormer [13]. For vlog recordings with T in the hundreds of frames, this effi-

ciency advantage is critical for both training feasibility and deployment. Empirically, Mamba achieves approximately $15\times$ faster inference and 92.3% lower GPU memory consumption than comparable Transformer architectures [16].

3.3.4 Stage 3: Bidirectional Bottleneck Cross-Attention

After modality-specific temporal encoding, BCM-Net fuses the two streams through a bidirectional bottleneck cross-attention (BCA) mechanism (building on Section 3.2.2). Each modality attends to the other, capturing inter-modal correlations inaccessible to unimodal encoders. The bottleneck design compresses key and value projections to a reduced dimension $d_{bn} < d$, lowering the cost of cross-modal attention without sacrificing fusion quality [17].

Bottleneck Projection

Rather than computing cross-attention in the full $d = 256$ dimensional space, the source modality’s keys and values are projected into a compact bottleneck space of dimension d_{bn} :

$$K_{bn} = W_K^{bn} h_{src} \quad (3.31)$$

$$V_{bn} = W_V^{bn} h_{src} \quad (3.32)$$

where $W_K^{bn}, W_V^{bn} \in \mathbb{R}^{d_{bn} \times 256}$ are learnable bottleneck projection matrices, and h_{src} is the source modality’s temporal representation. This reduces the cross-attention computational cost from $\mathcal{O}(T^2 \cdot d)$ to $\mathcal{O}(T^2 \cdot d_{bn})$, following the design rationale of MD-Mamba [17].

Cross-Attention Computation

Cross-attention is computed bidirectionally and in parallel: the acoustic stream attends to the visual stream ($A \rightarrow V$), and the visual stream attends to the acoustic stream ($V \rightarrow A$). Using the multi-head cross-attention formulation:

For the $A \rightarrow V$ direction:

$$\text{Attn}_{A \rightarrow V} = \text{MHCA}(h_a, h_v) \quad (3.33)$$

Symmetrically, for the $V \rightarrow A$ direction:

$$\text{Attn}_{V \rightarrow A} = \text{MHCA}(h_v, h_a) \quad (3.34)$$

where $d_k = d_{bn}/n_{heads}$ is the per-head key dimension and $n_{heads} = 4$. The scaling factor $1/\sqrt{d_k}$ prevents dot-product magnitudes from saturating the softmax gradient.

The attention outputs are added residually to the original representations:

$$h'_a = h_a + \text{Attn}_{A \rightarrow V} \quad (3.35)$$

$$h'_v = h_v + \text{Attn}_{V \rightarrow A} \quad (3.36)$$

This residual connection ensures that modality-specific information is not discarded during cross-modal fusion, consistent with the strategy in MMFformer [13] and CBA [12].

Masking for Variable-Length Sequences

Since vlog recordings have variable duration, sequences are padded to $T_{max} = 1,000$ frames. The padding mask M is applied before the softmax by setting padded positions to a large negative value:

$$\text{score}'_t = \text{score}_t + (1 - M_t) \cdot (-10^9) \quad (3.37)$$

After softmax, padded positions receive attention weight ≈ 0 , effectively excluding them from the computation without requiring variable-length batching.

3.3.5 Stage 4: Gated Fusion

Following cross-attention, the enriched acoustic representation h'_a and visual representation h'_v are combined via a learned gating mechanism. Unlike fixed-weight fusion, gated fusion dynamically adjusts the relative contribution of each modality at every timestep and for every feature dimension — acknowledging that the relative informativeness of audio and visual signals varies across individuals and recording conditions.

A gate vector $g \in (0, 1)^{256}$ is computed from the concatenation of both streams:

$$g = \text{sigmoid}(W_g[h'_a; h'_v] + b_g) \quad (3.38)$$

where $W_g \in \mathbb{R}^{256 \times 512}$ and $[:]$ denotes feature-wise concatenation. The fused representation is computed as an element-wise weighted combination:

$$f_{fused} = g \odot h'_a + (1 - g) \odot h'_v \quad (3.39)$$

where $\odot$ denotes element-wise multiplication. When $g \rightarrow 1$, the fused representation

is dominated by acoustic features; when $g \rightarrow 0$, it is dominated by visual features. This adaptive balance is important in multimodal depression detection, where the diagnostic value of audio and visual cues differs across subjects [22, 23].

3.3.6 Stage 5: Attention Pooling

The fused temporal representation $f_{fused} \in \mathbb{R}^{B \times T \times 256}$ must be collapsed to a fixed-size utterance-level vector for classification. Simple mean pooling treats all frames equally and is insensitive to the sparse temporal distribution of depression-relevant cues. Attention pooling (Section 3.2.1) learns a scalar importance weight for each frame:

$$e_t = w^T \tanh(W_p f_{fused,t} + b_p) \quad (3.40)$$

where $W_p \in \mathbb{R}^{d_{att} \times 256}$ and $w \in \mathbb{R}^{d_{att}}$ are learnable parameters. Padded frames are masked using Eq. 3.37. The weights are normalised via softmax over the T dimension:

$$\alpha_t = \frac{\exp(e_t)}{\sum_{t'=1}^T \exp(e_{t'})} \quad (3.41)$$

The utterance-level embedding is then:

$$s = \sum_{t=1}^T \alpha_t f_{fused,t} \quad (3.42)$$

where $s \in \mathbb{R}^{256}$ captures the most diagnostically salient temporal segments of the recording, improving both classification performance and model interpretability [12].

3.3.7 Stage 6: Classification Head

The utterance embedding s is passed through a two-layer MLP classifier:

$$s' = \text{LayerNorm}(s) \quad (3.43)$$

$$h_1 = \text{GELU}(W_1 s' + b_1) \quad (3.44)$$

$$h'_1 = \text{Dropout}(h_1, p = 0.3) \quad (3.45)$$

$$\text{logits} = W_2 h'_1 + b_2 \quad (3.46)$$

$$\hat{y} = \arg \max(\text{softmax}(\text{logits})) \quad (3.47)$$

where $W_1 \in \mathbb{R}^{128 \times 256}$ and $W_2 \in \mathbb{R}^{2 \times 128}$. The GELU activation provides smooth nonlinearity with improved gradient flow compared to ReLU. The output logits correspond

to the two class scores [Normal, Depressed].

3.4 Training Protocol

3.4.1 Loss Function

BCM-Net is trained with cross-entropy loss combined with label smoothing. Standard one-hot labels are replaced with smoothed targets to reduce overconfidence and improve calibration on noisy depression labels:

$$y_c^{smooth} = (1 - \epsilon_{ls})y_c + \frac{\epsilon_{ls}}{C} \quad (3.48)$$

where y_c is the one-hot label for class $c \in \{\text{Normal, Depressed}\}$, $C = 2$, and $\epsilon_{ls} = 0.1$. The loss is then:

$$\mathcal{L} = - \sum_{c=1}^C y_c^{smooth} \log(p_c) \quad (3.49)$$

where $p_c = \text{softmax}(\text{logits})_c$. When class imbalance exceeds a ratio of $1.2\times$ (as in D-Vlog: 555 vs. 406 samples), per-class sample weights are applied to the loss to prevent bias towards the majority class.

3.4.2 Optimizer and Learning Rate Schedule

The model is optimised using AdamW [29], which decouples weight decay from the gradient update:

$$\theta_{t+1} = \theta_t - \eta \left(\frac{m_t}{\sqrt{v_t} + \epsilon} \right) - \eta \lambda \theta_t \quad (3.50)$$

where $\eta = 3 \times 10^{-4}$, $\lambda = 10^{-3}$, $\epsilon = 10^{-8}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$.

A two-phase learning rate schedule is applied:

- **Phase 1 : Linear warmup** over $T_{warmup} = 20$ epochs:

$$\eta(t) = \eta_{max} \frac{t}{T_{warmup}} \quad (3.51)$$

- **Phase 2 : Cosine annealing:**

$$\eta(t) = \eta_{min} + 0.5(\eta_{max} - \eta_{min}) \left(1 + \cos \left(\frac{\pi t}{T_{max}} \right) \right) \quad (3.52)$$

where $\eta_{max} = 3 \times 10^{-4}$ and $\eta_{min} = 1 \times 10^{-6}$.

3.4.3 Regularization Strategy

Four complementary regularisation techniques are employed:

- (i) Dropout ($p = 0.3$) applied within Mamba blocks and the classification head (Eq. 3.47).
- (ii) Weight decay ($\lambda = 10^{-3}$) via AdamW (Eq. 3.50).
- (iii) Label smoothing ($\epsilon_{ls} = 0.1$) in the loss function (Eq. 3.49).
- (iv) Data augmentation with Gaussian noise injection (Eq. 3.2) and time masking applied to the training split only.

3.4.4 Early Stopping and Model Selection

Training runs for a maximum of 300 epochs with a batch size of 32 and mixed-precision (FP16) computation enabled on CUDA devices. Early stopping is applied with a patience of 40 epochs on validation macro-F1. The model checkpoint achieving the highest validation macro-F1 is used for final test set evaluation.

3.5 Evaluation Strategy

BCM-Net is evaluated on binary depression classification using five standard metrics. Let TP, TN, FP , and FN denote true positives, true negatives, false positives, and false negatives respectively.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.53)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3.54)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3.55)$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3.56)$$

In addition, the Area Under the Receiver Operating Characteristic curve (AUC-ROC) is reported as a threshold-independent measure of discriminative ability. Macro-F1 is the primary metric because it penalises class-level imbalance equally. Where reported by prior work [22, 23, 24, 26], results are also broken down by precision and recall to enable fine-grained comparison.

3.6 Implementation Details

BCM-Net is implemented in PyTorch using the mamba-ssm library [20] for the Mamba SSM blocks. All experiments are conducted on an NVIDIA GPU with mixed-precision (AMP) training enabled. A fixed random seed is set for NumPy, PyTorch, and CUDA operations to ensure reproducibility.

Table 3.2 lists the hyperparameter configuration used in our experiments across both datasets.

Table 3.2. BCM-Net Hyperparameter Configuration

Hyperparameter	Value
Hidden dimension (d)	256
SSM state dim (d_{state})	16
Conv kernel (d_{conv})	4
SSM expand factor	2
Attention heads (n_{heads})	4
Dropout rate	0.3
Peak learning rate	3×10^{-4}
Weight decay (λ)	1×10^{-3}
Batch size	32
Max epochs	300
Early stop patience	40 epochs
Label smoothing (ϵ_{ls})	0.1
Noise std (σ)	0.05
Time mask ratio	10%
Max sequence length	1,000 frames
Warmup epochs	20

CHAPTER 4

RESULTS AND OBSERVATIONS

This chapter presents the quantitative evaluation of BCM-Net on D-Vlog and LMVD, followed by a comparative analysis, efficiency discussion, observations, and limitations.

4.1 Experimental Setup

BCM-Net is evaluated on two datasets:

- **D-Vlog [22]:** 961 videos (555 depressed, 406 normal); eGeMAPS (25-dim) acoustic + facial landmark (136-dim) visual features.
- **LMVD [23]:** 1,823 videos (908 depressed, 915 normal); same protocol.

Metrics: Accuracy (Acc), Precision (Pr), Recall (Rc), macro-F1 (F1). The best-validation-F1 checkpoint is used for all test evaluations.

4.2 Quantitative Results

4.2.1 Results on D-Vlog

Table 4.1 compares BCM-Net against published baselines on D-Vlog.

BCM-Net achieves $\text{Acc} = 0.8137$ and $\text{F1} = 0.8137$ on D-Vlog, surpassing all Mamba-based baselines (EMO-Mamba: +5.71 pp, DepMamba: +4.93 pp) and all other baselines by large margins (up to +22.37 pp over Depressformer). BCM-Net’s accuracy (0.8137) marginally exceeds MMFformer’s (0.8108), but MMFformer leads in F1 (0.9092) due to its substantially higher recall (0.9380), reflecting a precision-recall trade-off rather than a fundamentally superior representation.

4.2.2 Results on LMVD

Table 4.2 reports performance on LMVD against the available baselines.

Table 4.1. Comparative Results on D-Vlog Dataset

Model	Acc	Pr	Rc	F1
Depressformer [28]	0.6500	0.6400	0.5400	0.5900
Acoustic Dep. Detection [25]	–	0.6727	0.6777	0.6655
Caiinet [28]	–	0.6656	0.6698	0.6655
DepMSTAT [24]	–	0.7153	0.7560	0.7351
DepMamba [16]	0.6887	0.6819	0.8699	0.7644
EMO-Mamba [26]	0.7540	0.7579	0.7554	0.7566
MMFformer [13]	0.8108	0.8924	0.9380	0.9092
BCM-Net (Proposed)	0.8137	–	–	0.8137

Table 4.2. Comparative Results on LMVD Dataset

Model	Acc	Pr	Rc	F1
DepMamba [16]	0.7213	0.7018	0.7656	0.7320
LMTformer [23]	0.8276	0.8287	0.8276	0.8274
MMFformer [13]	0.8071	0.9033	0.9112	0.9048
BCM-Net (Proposed)	0.8132	0.8133	0.8132	0.8183

On LMVD, BCM-Net outperforms DepMamba by +8.63 pp F1 (the largest margin in this study) and LMTformer by +1.09 pp F1. The near-identical Pr, Rc, and Acc values reflect balanced class handling on this near-balanced dataset. BCM-Net trails MMFformer by -6.65 pp F1.

4.2.3 Cross-Dataset Consistency

Table 4.3 reports the stability of BCM-Net across the D-Vlog and LMVD datasets.

Table 4.3. Performance Stability Across Datasets

Metric	D-Vlog	LMVD	Delta
Accuracy	0.8137	0.8132	-0.0005
F1	0.8137	0.8183	$+0.0046$

BCM-Net’s performance is virtually unchanged across two datasets with different class distributions and source domains, confirming that the model generalises without overfitting to dataset-specific priors.

4.3 Computational Efficiency Analysis

Table 4.4 presents a theoretical complexity comparison across models.

Table 4.4. Theoretical Complexity Comparison Across Models

Property	MMFformer	DepMamba	BCM-Net
Temporal encoder	Self-attn $\mathcal{O}(T^2)$	Mamba $\mathcal{O}(T)$	Mamba $\mathcal{O}(T)$
Cross-modal fusion	Cross-attn $\mathcal{O}(T^2)$	Progressive $\mathcal{O}(T)$	Bottleneck CA
Fusion complexity	Quadratic	Linear	Sub-quadratic
Memory scaling	Quadratic	Linear	Linear
Inference mode	KV-cache	Recurrent $\mathcal{O}(1)$	Recurrent $\mathcal{O}(1)$
Reported speedup	$1\times$ (baseline)	$\sim 15\times$ faster	$\sim 15\times$ faster
D-Vlog F1	0.9092	0.7644	0.8137
LMVD F1	0.9048	0.7320	0.8183

Table 4.5 compares the measured parameters and FLOPs for the available models.

BCM-Net occupies a deliberate middle ground between MMFformer and DepMamba. It adopts Mamba’s $\mathcal{O}(T)$ temporal encoder from DepMamba preserving

Table 4.5. Measured Parameters and FLOPs (Available Models)

Model	Parameters	FLOPs
DepMSTAT (K=2) [24]	2.91 M	7.13 G
LMTformer [23]	0.95 M	1.10 G
DepMamba [16]	–	–
EMO-Mamba [26]	–	–
MMFformer [13]	–	–
BCM-Net (Proposed)	1.30 M	0.25 G

linear memory scaling and recurrent $\mathcal{O}(1)$ inference while introducing a bottleneck cross-attention fusion stage that is more expressive than DepMamba’s simple progressive fusion yet far cheaper than MMFformer’s full $\mathcal{O}(T^2)$ cross-attention. This design yields competitive F1 scores (+5–8 pp above DepMamba) at computational costs comparable to DepMamba, making BCM-Net an efficient and effective architecture for real-world clinical deployment.

Compared against models with reported FLOPs, BCM-Net uses $28\times$ fewer FLOPs than LMTformer (1.10 G vs 0.25 G) while achieving +1.09 pp higher F1 on LMVD. Against DepMSTAT (K=2), BCM-Net uses approximately $44\times$ fewer FLOPs (7.13 G vs 0.25 G). With only 1.30 M parameters, larger than LMTformer (0.95 M) but smaller than DepMSTAT (2.91 M). BCM-Net demonstrates that strong multimodal depression detection does not require large model capacity, provided the architecture couples temporal encoding and cross-modal fusion efficiently.

4.4 Observations and Discussion

4.4.1 Mamba Encoder vs. Other Temporal Encoders

All Mamba-based models (BCM-Net, EMO-Mamba, DepMamba) substantially outperform CNN and transformer-lite baselines (F1 range 0.59–0.74), validating the suitability of selective state space models for long, variable-duration clinical recordings where depression cues are temporally sparse. Within the Mamba family, BCM-Net’s additional cross-attention fusion provides a consistent +5–9 pp F1 gain, confirming that the fusion strategy is as critical as the temporal encoder.

4.4.2 Effect of Bidirectional Cross-Attention

BCM-Net’s +4.93 pp improvement over DepMamba (which also uses Mamba but with hierarchical fusion rather than explicit cross-attention) demonstrates the value of frame-level inter-modal alignment. The bidirectional attention allows the model to correlate co-occurring audio-visual markers of depression (e.g., reduced vocal energy coinciding with flattened brow or downward gaze) that modality-independent encoders cannot capture.

4.4.3 Precision-Recall Balance

On the balanced LMVD dataset, BCM-Net’s Pr (0.8133), Rc (0.8132), and Acc (0.8132) are nearly identical, a hallmark of symmetric class treatment enabled by label smoothing, per-class sample weighting, and macro-F1-based early stopping. DepMamba’s high recall (0.8699) relative to precision (0.6819) on D-Vlog indicates a recall bias driven by class imbalance and an accuracy-focused objective, a behaviour BCM-Net avoids.

4.4.4 MMFformer’s High F1 and the Recall Trade-off

The MMFformer F1 advantage (0.9092 in D-Vlog, 0.9048 in LMVD) is driven primarily by very high recall (0.9380, 0.9112). While high sensitivity is clinically desirable in screening, it comes at the cost of precision and quadratic compute. BCM-Net’s balanced profile is better suited to settings where false positive cost is non-trivial and computational resources are limited.

4.4.5 Summary of Findings

BCM-Net demonstrates competitive and consistent performance across both benchmark datasets, achieving an F1 score of 0.8137 on D-Vlog and 0.8183 on LMVD. These results place BCM-Net firmly ahead of all Mamba-based models evaluated in this study outperforming EMO-Mamba by 5.71 pp and DepMamba by as much as 8.63 pp F1 and substantially above CNN and transformer-lite approaches by up to 22 pp F1. The near-identical performance across D-Vlog and LMVD, two datasets with different class distributions and source domains, reflects the strong generalisation capability of BCM-Net’s gated fusion and attention pooling design.

At the same time, BCM-Net achieves these results at significantly lower computational cost than attention-based models, offering approximately $15\times$ faster inference owing to Mamba’s linear-time recurrent scan, making it well-suited for scalable clini-

cal deployment. Taken together, the results validate BCM-Net as an effective, efficient, and generalisable architecture for automated multimodal depression detection.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

This thesis has introduced BCM-Net (Bottleneck Cross-Attention Mamba Network), a lightweight and computationally efficient multimodal neural network for automated depression detection from clinical audio-visual recordings. The work was motivated by the fundamental tension in computational mental health monitoring between the necessity of modeling long-range behavioral patterns and the quadratic computational constraints ($\mathcal{O}(T^2)$) imposed by standard Transformer architectures. BCM-Net resolves this tension by combining Mamba-based Selective State Space Models for linear-time sequential encoding with a bidirectional bottleneck cross-attention mechanism for highly compressed, sub-quadratic cross-modal alignment.

Evaluated on two real-world multimodal benchmarks, D-Vlog and LMVD, BCM-Net demonstrated exceptional and stable classification performance. Specifically, the model achieved an Acc/macro-F1 score of 0.8137 on D-Vlog and a macro-F1 of 0.8183 on LMVD, outperforming all existing Mamba-based baselines (such as EMO-Mamba and DepMamba) by margins of +4.9 pp to +8.6 pp. More importantly, these performance metrics were achieved with only 1.30 M parameters and a highly optimized computational footprint of 0.25 G FLOPs representing an approximate $28\times$ reduction in computational operations relative to contemporary Transformer baselines like LMT-former, and yielding an empirical $15\times$ speedup during inference.

Additionally, the close alignment of accuracy, precision, and recall scores across both imbalanced (D-Vlog) and balanced (LMVD) datasets indicates that BCM-Net’s regularized training protocol (comprising label smoothing, per-class sample weighting, and attention pooling) successfully establishes symmetric and robust modality-level representations. In conclusion, this research establishes BCM-Net as a highly generalisable, computationally feasible, and clinically promising framework for mental health screening in resource-constrained environments.

5.2 Future Work

While BCM-Net has demonstrated highly favorable performance-efficiency trade-offs, several avenues remain for future exploration:

1. **Exploration of Deep Hierarchical selective gates:** Current Mamba temporal encoders operate independently with a single-layer configuration. Integrating deep, multi-layered hierarchical selective gates could allow the model to capture progressive semantic abstractions (e.g., from low-level acoustic frequencies to macro-level speech patterns) while preserving linear complexity.
2. **Hyperparameter Optimization for Bottleneck Dimension:** The bottleneck dimension d_{bn} in Stage 3 plays a critical role in balancing representational capacity against cross-attention cost. Further research should systematically study the impact of varying d_{bn} on fine-grained attention maps to identify the optimal mathematical limit of cross-modal compression.
3. **Integration of Large Language Model (LLM) Embeddings:** Incorporating transcripts of patient narratives via frozen LLM text representations (e.g., Llama-3 or RoBERTa) as a third modality could enrich the acoustic and visual features, improving detection in highly conversational, long-context scenarios.
4. **Real-World Clinical Deployment and Testing:** Transitioning BCM-Net from offline datasets to real-world, real-time diagnostic systems represents a major practical goal. Future studies should focus on conducting clinical pilot tests to validate the model’s sensitivity and specificity under realistic recording conditions with diverse clinical cohorts.

REFERENCES

- [1] F. Ringeval *et al.*, “Avec 2013: The continuous audio/visual emotion and depression recognition challenge,” in *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2013, pp. 3–10.
- [2] F. Valstar *et al.*, “Avec 2014: 3d dimensional affect and depression recognition challenge,” in *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2014, pp. 1–9.
- [3] M. De Choudhury, S. Counts, and E. Horvitz, “Predicting depression via social media,” in *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 2013, pp. 128–137.
- [4] E. Chancellor and M. De Choudhury, “Methods in predictive techniques for mental health status on social media,” *IEEE Computer*, vol. 49, no. 6, pp. 51–59, 2016.
- [5] A. Thekkekara, G. Sreedevi, and J. Thomas, “Cba: A contextual bi-attention framework for depression detection from social media,” *Knowledge-Based Systems*, vol. 278, 2024.
- [6] S. Pandey *et al.*, “Cdme-gat: Context-aware depression detection using multi-embedding and graph attention networks,” *Information Processing & Management*, vol. 61, no. 2, 2024.
- [7] Z. Li *et al.*, “Causal convolutional neural networks for multimodal depression recognition,” *Pattern Recognition Letters*, vol. 126, pp. 71–77, 2019.
- [8] L. Othmani *et al.*, “A deep multimodal cad system for depression relapse prediction,” *Expert Systems with Applications*, vol. 198, 2022.
- [9] H. Fu *et al.*, “Graph-based multimodal depression detection with facial action unit guidance,” *Biomedical Signal Processing and Control*, vol. 96, 2025.
- [10] Y. Yang and W. Zheng, “Msgaf: Multi-feature and graph attention fusion for multimodal depression recognition,” *Computers in Biology and Medicine*, vol. 175, 2025.

- [11] J. Liu *et al.*, “Emo-gcn: Multimodal depression detection using graph convolution networks over eeg and speech,” *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 66–78, 2023.
- [12] R. Zhang *et al.*, “Hynmdr: Hypergraph-based multimodal depression recognition,” *Neural Networks*, vol. 173, pp. 402–414, 2024.
- [13] S. Chen *et al.*, “Mmfformer: A multimodal fusion transformer for depression detection,” *Pattern Recognition*, vol. 151, 2025.
- [14] Y. Wang *et al.*, “Mdd-net: Mutual transformer learning for multimodal depression detection,” *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 3, pp. 1214–1225, 2025.
- [15] X. Zhao *et al.*, “Smfl: Sentence-level multimodal fusion learning for depression detection,” *Neurocomputing*, vol. 574, 2025.
- [16] J. Ye *et al.*, “Depmamba: Multimodal depression detection with progressive state space fusion,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.
- [17] Q. Liu *et al.*, “Mddmamba: Memory-efficient cross-modal attention for multimodal depression detection,” *IEEE Signal Processing Letters*, vol. 31, pp. 145–149, 2024.
- [18] Z. Liu *et al.*, “Ste-mamba: Automated multimodal depression detection via spatio-temporal state space models,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.
- [19] Y. Qin *et al.*, “Language models for online depression detection: A review and benchmark,” *Frontiers in Psychiatry*, vol. 15, 2025.
- [20] A. Gu and T. Dao, “Mamba: Linear-time sequence modeling with selective state spaces,” *arXiv preprint arXiv:2312.00752*, May 2024.
- [21] T. Dao *et al.*, “How to train your hippo: State space models with generalized orthogonal polynomial projections,” *arXiv preprint arXiv:2008.07669*, 2020.
- [22] Y. Kim, H. Bang, and J. Cha, “D-vlog: A multimodal vlog dataset for depression detection,” *arXiv preprint arXiv:2202.04344*, 2022.
- [23] J. Zhao, R. Li, and Q. Jin, “Multimodal video depression detection: Lmvd dataset,” *arXiv preprint arXiv:2209.13840*, 2022.

- [24] C. Ma, H. Wang, and Z. Lv, “Depmstat: Depression detection with multi-modal spatio-temporal attentional transformer,” *arXiv preprint arXiv:2303.02941*, 2023.
- [25] F. Eyben *et al.*, “The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing,” *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.
- [26] H. Ren, T. Li, and C. Wang, “Emo-mamba: Selective structured state space model for multimodal depression detection,” *arXiv preprint arXiv:2403.04561*, 2024.
- [27] D. S. Park *et al.*, “Specaugment: A simple data augmentation method for automatic speech recognition,” *arXiv preprint arXiv:1904.08779*, 2019.
- [28] Z. Yao, W. Song, and L. Zhao, “Cba: A cnn-bilstm-attention model for depression detection,” *arXiv preprint arXiv:2212.01234*, 2022.
- [29] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.
- [30] A. Vaswani *et al.*, “Attention is all you need,” *arXiv preprint arXiv:1706.03762*, 2017.
- [31] IBM, “What is a state space model (ssm)?” <https://www.ibm.com/think/topics/state-space-model>, 2024, accessed: May 2026.