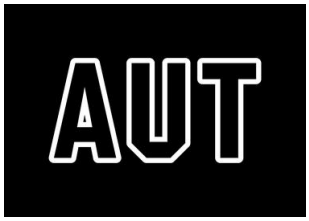

MLGA - A Modality-Level Graph Attention Architecture for Multimodal Depression Detection

Malika Malika

MASTER'S THESIS

A research component submitted to Auckland University of Technology in fulfilment of the requirements for the degree of

Masters of Philosophy (M.Phil)

Supervisor: Dr. Sira Yongchareon

School of Engineering, Computer and Mathematical Sciences

Auckland, September 2025

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgements), nor used artificial intelligence tools or generative artificial intelligence tools (unless it is clearly stated, and referenced, along with the purpose of use), nor material which to a substantial extent has been submitted for the award of any other degree or diploma from a university or other institution of higher learning.

Date 27-01-2026 **Signature** Malika

Acknowledgements

I would like to sincerely thank my supervisor, Dr. Sira Yongchareon, for his invaluable guidance, expertise, and encouragement throughout the course of this research. His constructive feedback and steady support were central to shaping my ideas and bringing this thesis to fruition.

I am also deeply appreciative of the academic staff and my peers at the Auckland University of Technology, whose insightful discussions, collaboration, and motivation greatly enriched my learning journey.

My heartfelt gratitude extends to my family and friends, whose constant support, patience, and encouragement have been a source of strength and inspiration during this endeavor. Their unwavering belief in me made this achievement possible.

Finally, I wish to acknowledge the wider research community, whose contributions, open-source resources, and scholarly work provided a foundation upon which this study has been built.

Auckland, September 2025

Malika Malika

Abstract

Depression is one of the most pressing global health challenges, affecting millions of individuals and placing significant strain on healthcare systems. Early and accurate detection is critical for timely intervention and improving patient outcomes. Traditional diagnostic methods, which rely heavily on clinical interviews and self-reports, are often resource intensive, subjective, and limited in scalability. To address these limitations, this study presents an architectural investigation of Modality-Level Graph Attention (MLGA), a deep learning framework for multimodal fusion in depression detection.

The proposed architecture integrates textual embeddings from ClinicalBERT, visual representations from VGG-PCA, and facial behavioral descriptors from OpenFace. These modalities are fused through a modality-level Graph Attention Network (GAT) that explicitly models inter-modality relationships, while a temporal module captures dynamic behavioral patterns over time. To enhance robustness, the framework incorporates Gaussian noise injection, L2 normalization, and modality dropout, thereby encouraging resilience to noise and missing inputs.

A comprehensive evaluation was conducted on three benchmark datasets: E-DAIC-WOZ, EATD-Corpus, and D-Vlog. Across these datasets, MLGA achieved competitive performance and, on the E-DAIC benchmark in particular, surpassed several unimodal and late-fusion baselines in terms of precision, recall, F1-score, and ROC-AUC, demonstrating the effectiveness of graph-based multimodal integration under the studied conditions.

The results highlight the importance of modeling both intra-modality features and cross-modality dependencies within a unified fusion architecture. Rather than proposing a fully deployable clinical tool, this study advances the field of affective computing by systematically analysing a modality-level graph attention design that is computationally moderate and interpretable, and by quantifying its behaviour across heterogeneous datasets. Future directions include expanding modality coverage, applying domain adaptation for cross-cultural and cross-setting generalization, and enhancing interpretability using advanced explainable AI techniques.

Keywords: Multimodal Depression Detection, ClinicalBERT, VGG, OpenFace, Graph Attention Networks, Temporal Modeling, Explainable AI

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
AU	Action Unit (facial muscle movement)
AUC	Area Under the ROC Curve
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
DAIC-WOZ	Distress Analysis Interview Corpus – Wizard of Oz
E-DAIC	Extended Distress Analysis Interview Corpus
EATD	Emotional Audio-Textual Depression Corpus
EHR	Electronic Health Record
F1	Harmonic mean of Precision and Recall
GAT	Graph Attention Network
GNN	Graph Neural Network
GRU	Gated Recurrent Unit
LSTM	Long Short-Term Memory
ML	Machine Learning
MLGA	Modality-Level Graph Attention (this thesis)
MLP	Multi-Layer Perceptron
NLP	Natural Language Processing
PCA	Principal Component Analysis
PHQ-8/PHQ-9	Patient Health Questionnaire (8-item / 9-item)
PR	Precision–Recall
ROC	Receiver Operating Characteristic
SHAP	SHapley Additive exPlanations
SOTA	State of the Art
SVM	Support Vector Machine
TGN	Temporal Graph Network
ViT	Vision Transformer
VGG	Visual Geometry Group (CNN architecture)
WHO	World Health Organization

Chapter 1

Introduction

Mental health underpins the well-being and productivity of societies; however, reliably recognizing depressive disorders in everyday contexts remains a persistent challenge across busy clinics, remote communities, and digital interactions. Depression is pervasive and burdensome, ranking among the leading contributors to global disability and public health costs [5, 6, 7]. Symptoms emerge gradually and include unevenly lowered mood, anhedonia, slowed speech, and reduced expressivity, often shaped by context and stigma, which delays help-seeking and obscures early warning signs.

Conventional assessments, structured clinical interviews, and self-report scales, such as the PHQ-9 and BDI, are indispensable in practice [8, 9]. However, they are time-intensive and subjective, and access to trained clinicians remains uneven, particularly in low-resource or geographically isolated settings [10]. Consequently, subthreshold or variable patterns are frequently overlooked, precisely when timely intervention would be most advantageous.

Outside the clinic, everyday digital traces carry rich and affective information. Linguistic choices in text (e.g., first-person singular, negative affect) [11, 12], facial dynamics and gaze patterns [13, 14], and prosodic contours in speech (e.g., flattened pitch, reduced variability) [15] correlate with depressive states. Crucially, no single channel suffices; integrating complementary modalities yields a more ecologically valid assessment than analyzing any one in isolation [16, 17, 18, 19].

Recent advances have made such integration practical at scale. Transformer encoders (e.g., BERT, ClinicalBERT) enable nuanced language understanding in interviews and narratives [1, 2], while responsibly designed digital mental health platforms and conversational agents demonstrate how AI can extend its reach and support continuous monitoring when implemented with appropriate safeguards [20, 21, 22].

Against this backdrop, this study proposes an interpretable, modality-level multimodal framework that fuses text, facial, and visual cues to support earlier, more consistent depression screening under real-world constraints. The design emphasizes calibrated decision-making, robustness to missing or degraded inputs, and computational efficiency at the fusion stage, with an eye toward eventual use in telehealth and mobile applications subject to further optimization and validation. The next sections provide a detailed motivation for this problem, justify the modelling choices, outline the research contributions, and map the structure of this study.

1.1 Motivation

The practical question is not merely whether AI can detect depression, but whether it can do so early, fairly, and reliably in the settings where care actually happens, such as telehealth consultations, mobile interfaces, and diverse community contexts. This perspective shifts the objective from pure accuracy on benchmark datasets to clinical utility: models must surface actionable risk information in time to change outcomes, behave consistently across populations, and operate under common constraints, such as noisy inputs, limited compute, and variable modality availability. Consequently, the design goal is a system that is accurate yet efficient, explainable yet data-driven, and resilient to missing or degraded input data.

Clinical and societal imperative

Earlier recognition of depressive symptoms is linked to improved outcomes, reduced relapse, and more efficient use of services [23]. However, in practice, cases are often identified late, after distress has escalated or functioning has deteriorated. Traditional pathways, clinician-administered interviews, and self-report measures remain indispensable, but they are constrained by time, access, and variability in administration [8, 9, 10]. A scalable, AI-assisted screen that highlights high-risk individuals for clinician review can narrow this gap, especially in systems with long waitlists or limited specialist coverage [10]. The aim is to complement not replace clinical judgment: automated triage should make early signals harder to miss and follow-up more targeted.

Economic and service delivery pressures

Depression imposes substantial costs on employers, health systems, and families through absenteeism, presenteeism, and recurrent health service utilization [24]. If screening tools can identify likely cases earlier and more consistently, interventions can be offered sooner, reducing avoidable downstream utilization and mitigating productivity losses. This requires models that are not only discriminative but also stable and calibrated to operate at thresholds that reflect service capacity and clinical priorities (e.g., prioritizing recall for safety vs. precision for resource stewardship).

Access, equity, and acceptability

Access to trained clinicians, clinic proximity, language proficiency, and willingness to disclose symptoms are unevenly distributed, creating structural barriers for underserved groups [10]. Digital screening can extend reach beyond clinic walls, but only if methods generalize across languages, cultures, and capture conditions and are implemented with transparency and con-

sent [22]. This motivates designs that are robust to distribution shifts and explicit about uncertainty, with reporting practices that make trade-offs visible to stakeholders.

Opportunity in everyday multimodal signals

Outside the clinic, routine digital interactions encode complementary affective cues: language use (e.g., negative affect, first-person singular) [11, 12], facial dynamics and gaze patterns [13, 14], and prosodic contours in speech (e.g., flattened pitch, reduced variability) [15]. No single channel is sufficient; the signals are often weak, intermittent, or occluded. Integrating modalities yields a more ecologically valid and resilient assessment than analyzing any one modality in isolation [16, 17, 18, 19]. The motivation here is to aggregate complementary evidence while controlling for spurious correlations introduced by noise, topics, or context.

Computational efficiency and deployability

For telehealth and mobile use, a model's practical value depends on computation and latency as much as accuracy. Large cross-modal transformers may be difficult to operate in real time or on devices. The framework used in this study is therefore compact at the fusion stage by design: 512-dimensional projections per modality, a single modality-level Graph Attention layer over a three-node graph (Text/Face/Visual), a 128-dimensional fused bottleneck, and a shallow classifier with calibrated thresholding. With pre-computed embeddings, the runtime cost is dominated by this small fusion head, which suggests that the architecture may be compatible with constrained environments when lightweight encoders or cached embeddings are available. Chapter 5 quantifies the parameters, memory, and latency to characterise the computational profile of the architecture, rather than to claim immediate deployability.

Taken together, these considerations motivate a system that: (i) integrates heterogeneous cues to capture subtle, dispersed indicators; (ii) is robust to noise and missing inputs; (iii) exposes interpretable signals at the modality level to support clinical scrutiny; and (iv) is efficient enough for real-world use. The following chapters translate these goals into concrete modelling choices, evaluation protocols aligned with clinical trade-offs, and a deployment-oriented analysis of efficiency and reliability.

1.2 Research Justification

Depression assessment in routine practice relies on clinician-led interviews and self-report instruments, such as the PHQ-9 and Beck's Depression Inventory (BDI) [8, 9]. While indispensable, these tools are vulnerable to subjectivity, stigma-driven under-reporting, variability in administration, and constraints on time and access, which can cause subtle or fluctuating

symptom patterns to be missed [10]. This motivates the need for computational support that can consistently surface early signals and complement clinical judgment rather than replace it [20, 21]. Simultaneously, any technological aid must function where care is actually provided, such as telehealth clinics, mobile devices, and culturally diverse communities, without adding an undue operational burden.

A purely unimodal pipeline (text-only, audio-only, or vision-only) can capture channel-specific markers; however, it degrades when the chosen channel is noisy or absent [15, 12, 25, 26]. Even within multimodal research, widely used integration strategies have drawbacks: simple early fusion ignores structure and scale disparities between modalities, and late fusion combines decisions after the fact, discarding cross-modal synergies that matter precisely when one channel weakens. Meanwhile, very deep cross-modal transformers often exceed the latency and memory budgets of real deployments, limiting their use outside laboratory conditions [27]. Therefore, the approach adopted in this study was selected to address these practical constraints while remaining accurate and robust, and to explore a fusion design that is computationally feasible under typical telehealth-like resource budgets.

The first requirement is sensitivity to fine-grained and transient indicators that may be weak or sparsely expressed in any single channel. To avoid overfitting to isolated frames or utterances while still capturing early cues, the method computes segment-level embeddings from ClinicalBERT for text, OpenFace for facial behavior, and VGG with PCA for visual appearance, and then models temporal dynamics across segments using a lightweight GRU [26, 1, 2]. Segment granularity preserves brief events, temporal aggregation improves sensitivity to dispersed evidence, and stabilizes decisions against momentary noise. This directly addresses the risk that fleeting indicators are lost in per-frame or per-utterance analysis.

The second requirement is resilience to nuisance variation and recording artifacts that confound features in the wild [25]. The system normalizes scale and variance via compact linear projections (to 512d) and PCA; during training, it uses modality dropout so that the model learns to remain predictive when inputs degrade; at inference, it handles absent channels by masking nodes and edges in the fusion graph [28]. These mechanisms jointly reduce false positives caused by lighting, pose, topic shifts, or preprocessing differences, and prevent brittle failure when inputs are incomplete, which are typical conditions in telehealth and mobile capture.

Third, because performance can drop across real-world variability, including languages, cultures, and capture conditions under-represented in training evaluation, the architecture and experimental protocol are structured to probe generalization directly. We therefore benchmark across three complementary corpora with differing natural modality availability: E-DAIC (Text+Visual+Face) [29], EATD (Text) [30], and D-Vlog (Visual) [31]. Using datasets that reflect different capture contexts and modality constraints stresses robustness to distribution

shift and ensures that the method is exercised under conditions that approximate likely usage scenarios, without claiming full deployment readiness.

The fourth design goal is to capture cross-modal dependencies rather than assuming independence. Simple concatenation treats modalities as unrelated features, and late fusion forfeits the ability to share information when a channel is weak. Instead, the system constructs a modality graph over {Text, Face, Visual} and applies a single Graph Attention (GAT) layer to learn directed, context-dependent influences among modalities before temporal modelling [32]. This explicitly encodes the inter-modality structure, enabling the model to reweight channels over time and draw on the most reliable source as conditions vary across an interview.

Fifth, interpretability and trust are essential for clinical applications. While full mechanistic transparency is unrealistic for modern classifiers, the design provides modality-level attention coefficients as an attribution signal indicating which channel contributed most under given conditions, and it uses validation-calibrated thresholds to report operating points that reflect clinically meaningful precision–recall trade-offs [33]. These artifacts support clinician review, documentation, and contestability without overstating the degree of interpretability.

Finally, the approach must respect the computational constraints that are typical of telehealth and mobile scenarios, without over-claiming production readiness. The fusion stack is compact by construction, comprising linear projections, a single GAT over three nodes, a 128-d bottleneck, and a single-layer GRU. With precomputed embeddings, the online cost is dominated by this small fusion–temporal–classifier block, aiming to respect real-time latency and memory budgets while retaining the benefits of multimodal integration. Chapter 5 quantifies the parameter counts, memory footprint, and latency, and contrasts these with deeper fusion baselines to characterise the computational profile of the architecture, rather than to claim immediate deployability.

The empirical protocol mirrors realistic usage. Each dataset was exercised under its natural modality constraints, class imbalance was handled explicitly, and results were reported with calibrated thresholds, ROC/PR curves, and full confusion matrices to ensure that precision–recall trade-offs were transparent and reproducible. Taken together, the architectural choices and evaluation design address the practical obstacles above, enhancing sensitivity to subtle signals, improving robustness to noise and missing inputs, capturing cross-modal structure, supporting clinician trust, and approximating resource budgets typical of telehealth-like settings. The central contribution of this thesis is thus an empirical and architectural study of modality-level graph attention for multimodal fusion in depression detection [16, 17, 33].

1.3 Significance and Contributions of this Research

This study primarily contributes an architectural investigation of modality-level graph attention for multimodal fusion in depression detection, with the following specific contributions to the domain of multimodal depression detection:

- **Innovative Methodology:** We developed a modality-level fusion architecture that projects text, facial, and visual embeddings into compact 512-dimensional spaces and integrates them using a single Graph Attention (GAT) layer. The fused 128-dimensional representation is then modelled temporally with a GRU and classified with a calibrated shallow head. This design explicitly captures cross-modal relationships while remaining computationally moderate under the evaluated settings and is proposed as a candidate architectural template for future real-world applications, rather than as a fully engineered deployment-ready system.

- **Enhanced Sensitivity to Subtle Indicators:** By aggregating complementary linguistic, facial, and visual cues, the proposed approach improves sensitivity to early or weak signals that may be under-expressed in any single modality. The fusion mechanism allows stronger channels to compensate when others are sparse or attenuated, thereby supporting earlier and more reliable screening within the constraints of the available data.

- **Increased Robustness and Reliability:** The model is designed to operate under realistic constraints, where the inputs may be noisy, incomplete, or privacy-restricted. Training-time modality dropout and inference-time masking of nodes/edges in the modality graph yield graceful degradation when one or more modalities are missing or degraded, reducing brittle failure modes in telehealth-like and in-the-wild settings [28].

- **Modelling Cross-Modal and Temporal Dependencies:** The modality-level GAT explicitly encodes the relationships among the Text, Face, and Visual streams, whereas the GRU captures the intra-session temporal dynamics without incurring the overhead of deep cross-modal transformers. This combination improves the representational fidelity for fluctuating symptom expression within interviews [27] and provides a concrete comparative point against traditional early- and late-fusion baselines.

- **Characterisation of Computational Efficiency:** The fusion+temporal+classifier stack is compact by construction (linear projections, single GAT layer, 128-d bottleneck, shallow classifier). When embeddings are pre-computed, the online inference cost is dominated by this small stack. Chapter 5 reports parameter counts, memory footprint, and latency to characterise the computational profile of the architecture and to illustrate that it may be compatible with resource-constrained or mobile/telehealth scenarios, subject to further

optimisation, validation, and systems integration. The reported parameter counts therefore describe the fusion and classification stack on top of precomputed embeddings; in any real-time setting that processes raw inputs, the computational cost of the full ClinicalBERT and VGG-16 backbones would also need to be considered, and lighter encoders or compression strategies would likely be required for mobile devices.

- **Reproducible, Clinically-Oriented Evaluation:** We aligned modality usage with dataset reality: E-DAIC (Text+Visual+Face), EATD-Corpus (Text), and D-Vlog (Visual) to ensure practice-relevant benchmarking. The study employed imbalance-aware training, validation-tuned thresholds, and full confusion matrices, making precision–recall trade-offs transparent and facilitating clinical interpretation and replication. This evaluation provides an empirical basis for understanding the strengths and limitations of modality-level graph attention fusion across heterogeneous datasets.

1.4 Research Questions

Motivated by the gaps and practical constraints identified in Chapter 1 and the multimodal depression detection literature reviewed in Chapter 2, this thesis is guided by the following research questions:

- **RQ1:** Does a modality-level graph attention mechanism improve depression-detection performance compared with unimodal models and simple feature concatenation or late-fusion baselines on the E-DAIC-WOZ benchmark?

- **RQ2:** What are the relative contributions of textual, visual, and facial modalities, and of temporal modelling and regularisation components, to the overall performance of the proposed framework?

- **RQ3:** How robust is the proposed modality-level graph attention architecture under dataset and modality shifts, as assessed by transferring it from E-DAIC-WOZ to the EATD-Corpus (text-only, language-mismatched) and D-Vlog (visual-only, in-the-wild) benchmarks?

RQ1 focuses on the core architectural contribution of this work, namely the use of modality-level Graph Attention Networks (GAT) for multimodal fusion. RQ2 probes the importance of individual modalities and architectural components through ablation analysis and attention-weight inspection. RQ3 examines robustness and generalization under demographic, linguistic, and capture-condition shifts by evaluating the model on complementary datasets with differing modality availability and domain characteristics.

1.5 Thesis Structure

In Chapter 2, we review unimodal and multimodal depression detection methods, including attention and graph-based fusion, and identify gaps in robustness, generalization, and interpretability [16, 17]. Chapter 3 presents the proposed architecture: 512-d projections per modality, a single modality-level GAT, a 128-d fused bottleneck, GRU temporal modeling, and a calibrated classifier [32]. Chapter 4 details datasets, preprocessing, and feature extraction (ClinicalBERT, OpenFace, VGG/PCA) [1, 2]. Chapter 5 reports the experimental setup, metrics, and results of the ablation studies. Chapter 6 discusses the implications and limitations, and Chapter 7 concludes with directions for future work.

1.6 Summary

This chapter established the need for scalable, objective depression screening and motivated a lightweight, interpretable multimodal approach that fuses text, facial, and visual signals. We framed the practical constraints (access, equity, robustness, and deployability), articulated why unimodal and naïve fusion pipelines fall short, and justified the modality-level Graph Attention design with temporal modelling and calibrated decision thresholds. We outlined contributions that target robustness to missing modalities, clinically oriented reporting, and computational efficiency that is intended to be compatible with telehealth and mobile contexts, subject to further optimisation and validation. The next section reviews related work and positions our method within the broader landscape of unimodal and multimodal depression detection [16, 17, 22].

Chapter 2

Literature Review

Over the past decade, computational approaches for detecting depression have significantly progressed. This growth is largely driven by the global prevalence of mood disorders and the limitations of conventional clinical assessments [5, 20]. Research in this area spans both machine learning and deep learning techniques, applied to diverse modalities such as language, speech, and visual signals, either individually or in multimodal combinations [16, 21, 15, 34]. This chapter systematically reviews this field. It begins by outlining the general framing of depression detection (Section 2.1), followed by the characteristics and challenges of the task (Section 2.2), and then traces the methodological evolution from manual assessments to multimodal deep learning approaches incorporating attention and graph-based fusion techniques (Section 2.3). Recent studies have also emphasized issues of fairness, interpretability, and computational scalability [35, 36, 37].

2.1 Introduction to Depression Detection

In the context of this study, depression detection refers to the computational estimation of depressive symptoms based on observable, behavioral, and linguistic signals. These signals are typically captured in structured interviews, telehealth consultations, and naturalistic settings, such as social media interactions. The primary objective of such systems is not to replace professional clinical judgement but to serve as an auxiliary tool that can help identify at-risk individuals and support ongoing monitoring, particularly in contexts where access to trained mental health specialists is limited [5].

The input evidence for computational detection is generally derived from three sources. Textual data include spoken transcripts from interviews, self-reported questionnaires, or user-generated content such as social media posts, all of which can reveal linguistic markers associated with depression. Acoustic data capture features of the voice, such as prosody, intonation, energy, and voice quality, which often reflect psychomotor changes and emotional states. Visual and facial behavior includes observable cues such as facial expressivity, gaze direction, eye contact, and head movements, which provide insight into affective engagement and nonverbal communication patterns [16, 21]. These modalities are complementary, as each channel encodes

different aspects of behavior relevant to depressive symptomatology.

Detection can be performed at varying levels of granularity depending on the research design. Some systems focus on short segments, such as utterances or specific responses, to detect local markers of depression, whereas others operate at the full-session level, summarizing information across an entire interview or video. Similarly, studies have been conducted across a range of capture environments, including highly structured clinical or telehealth settings, controlled laboratory protocols, and naturalistic "in-the-wild" contexts, such as online videos or social media streams [15, 34]. Each setting imposes different constraints on signal quality, ecological validity, and generalizability.

Automated depression detection systems must address several persistent challenges across all contexts. Recordings are often noisy, with background interference in the audio or occlusions in the video. Cultural and demographic diversity complicates generalization, and depressive markers themselves tend to occur sparsely, often triggered only under certain conversational topics. Beyond technical robustness, systems must also be designed with transparency, interpretability, and practicality in mind so that outputs can be trusted and integrated into real-world workflows without the risk of misinterpretation or harm [36, 35].

2.2 Nature and Classification of Depression Detection

Depression detection is inherently a multimodal task because human communication conveys psychological states through multiple complementary channels. Linguistic choices, such as word use and discourse structure, reflect underlying affective and cognitive processes. Speech, through rhythm, tone, and energy, encodes psychomotor changes that are often altered in individuals with depression. Visual behavior, including facial expressions, gaze direction, and posture, provides further evidence of engagement and emotional responsiveness. Together, these three modalities: text, audio, and visual, form a rich but complex information space for computational models [16, 21, 34]. However, depressive cues are not always consistent. They often appear intermittently, and their manifestation depends heavily on the conversational context or external conditions. This irregularity necessitates systems that can selectively attend to salient cues while remaining robust to noise, missing data, or degraded signals.

To address this complexity, prior research commonly classifies depression detection studies into four practical dimensions. These dimensions influence how data are collected, annotated, and modelled, and they help clarify the assumptions behind methodological choices.

1. **Modality configuration** Systems may rely on a single modality, such as text-only, audio-only, or visual-only, or they may combine multiple channels in multimodal configurations. Multimodal approaches generally achieve higher performance by leveraging comple-

mentary signals; however, they introduce additional challenges in synchronization and fusion [16].

2. **Capture setting** The environment in which data are gathered significantly affects both the quality and generalizability. Clinical and telehealth interviews provide structured diagnostic settings, laboratory protocols enable controlled capture conditions, and "in-the-wild" sources, such as social media or personal video blogs, capture more naturalistic behavior at the cost of greater variability and noise [21, 35].

3. **Temporal scope** Some approaches adopt a point-in-time perspective and estimate depressive symptoms in a single session. Others use a longitudinal framework, monitoring individuals across multiple sessions to capture the evolution of symptoms, relapse patterns, or gradual changes. This distinction has methodological implications because longitudinal studies typically require temporal models, such as recurrent networks or temporal graph modules [15].

4. **Decision granularity** Models differ in whether they output predictions at the utterance or segment level (with later aggregation) or generate direct session-level classification. Fine-grained approaches allow the localization of depressive cues and greater interpretability, whereas session-level predictions are simpler but may obscure the dynamics of symptom expression [36].

These four axes provide a useful taxonomy for depression detection research. They clarify assumptions about data quality, highlight the challenges of alignment and generalization, and guide the selection of features, encoders and fusion strategies. This structured perspective also helps identify the trade-offs between ecological validity and modelling complexity, which are critical considerations when designing systems intended for clinical or real-world deployment.

Figure 2.1 summarizes this taxonomy, highlighting how modality configuration, capture setting, temporal scope, and decision granularity are related to the central task of depression detection.

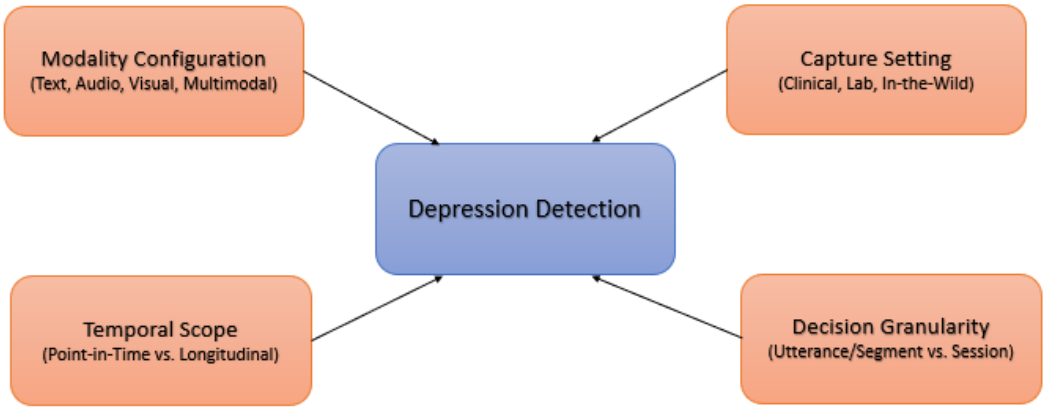

Figure 2.1: Taxonomy of depression detection along four practical axes, modality configuration, capture setting, temporal scope, and decision granularity, illustrating how each dimension informs data preparation and modelling choices. Author-created schematic

2.2.1 Challenges in Depression Detection

Despite significant progress, depression detection remains a complex research area because practical deployments must contend with various real-world challenges.

- **Noisy and incomplete inputs** Inputs across modalities are often missing or degraded due to noise. Microphones may capture background noise or clip speech, cameras may lose track of a face or operate in poor lighting, transcripts may contain errors, and social media text is frequently short, informal, or ambiguous. These limitations make it essential to design systems that remain functional even when some evidence is unreliable or absent [16].

- **Alignment across modalities** Text, audio, and visual signals evolve over different timescales. Spoken words unfold quickly, prosodic features such as pitch or rhythm span longer intervals, and facial expressions may change gradually. If these streams are poorly aligned, the benefits of multimodal fusion are reduced, and contradictory cues may arise [16].

- **Sparse and context-dependent cues** Depressive indicators rarely occur continuously. Signs such as monotone prosody, lowered facial expressivity, or negative self-focused language may appear only briefly and in response to specific prompts. Detecting these transient signals requires temporal models that can focus on informative regions while aggregating across longer interactions [11, 14].

- **Label quality and protocol differences** Ground-truth annotations are usually derived from self-report instruments such as the PHQ-9 or BDI, or from structured clinical interviews. However, thresholds, administration conditions, and rater variability differ across datasets, introducing inconsistencies that complicate cross-corpus comparisons [30].

- **Imbalance and limited sample size** In most datasets, participants with clinically significant depression were outnumbered by controls. This imbalance makes optimization more difficult and can bias the evaluation if not carefully addressed. Therefore, transparent reporting of decision thresholds and multiple metrics is required to ensure a fair assessment [36].

- **Domain shift** Models trained on one dataset may not generalize well to others because of differences in language, cultural norms, recording devices, or interview protocols. Without explicit adaptation strategies, domain shifts can severely degrade performance [30, 36].

- **Interpretability and calibration** For clinical use, systems must not only be accurate but also provide interpretable and reliable outputs. Clinicians must understand the rationale behind a decision, and probability scores must be properly calibrated to ensure that risk estimates are meaningful. Opaque or overconfident predictions limit adoption [38].

- **Privacy, consent, and fairness** Depression datasets often include personal narratives and identifiable facial behaviors, raising serious privacy concerns. Furthermore, if datasets under-represent certain groups, models risk producing biased predictions and exacerbating care disparities. Therefore, responsible data governance and fairness-aware evaluation are critical [22, 39, 40, 41].

- **Operational constraints** Practical applications, such as telehealth consultations or mobile-based monitoring, require systems that operate under the strict constraints of latency, memory, and power. This creates a need for efficient encoders and lightweight fusion strategies that can function reliably in real-world deployments [37].

2.2.2 The Critical Importance of Depression Detection

Early and reliable detection of depression is critical for both individuals and healthcare systems. Depression is often underdiagnosed and untreated, particularly in regions where access to qualified mental health professionals is limited. Computational approaches can help address this gap by offering scalable tools that extend the reach of the existing services. By identifying individuals at risk earlier, these systems can enable proactive referrals, continuous monitoring, and relapse prevention. Such interventions not only alleviate the personal burden of symptoms but also reduce the wider societal and economic costs associated with untreated depression [5].

Digital methods also provide opportunities for continuous and asynchronous screenings. For example, automated triage systems embedded in telehealth platforms can monitor patients between clinical visits, flagging concerning changes in behavior or moods. Importantly, these systems are designed to complement clinical expertise rather than replace it, supporting practitioners with additional insights derived from multimodal data.

Several requirements must be met for computational detection systems to be clinically credible. They must provide transparent reasoning, offering interpretable outputs that clinicians can trust; they must ensure calibrated predictions, where probability estimates meaningfully reflect true risk; and they must include documented limitations, clarifying the contexts in which performance may be degraded. Furthermore, validation must extend across diverse demographic groups, cultural contexts, and capture conditions to ensure fairness and avoid unintended harm. Without these safeguards, even highly accurate systems risk producing biased or misleading outcomes, ultimately undermining their potential benefits.

2.3 Evolution of Depression Detection Methods

2.3.1 Manual and Instrument-Based Assessment

Traditional approaches to depression assessment remain centered on clinician-led interviews and validated self-report questionnaires, such as the Patient Health Questionnaire (PHQ-9) and Beck Depression Inventory (BDI). These methods are considered the clinical gold standard because they provide high validity and interpretability and remain the source of ground-truth labels for most computational research [5]. However, they have important limitations: structured interviews are time-intensive, access to trained professionals is limited in many regions, and administration protocols vary across raters and contexts, leading to inconsistencies in the results. Therefore, computational methods are positioned as complementary tools rather than replacements, offering opportunities to increase reach, enable consistent and repeatable screening, and support early identification in resource-constrained environments [20].

2.3.2 Single-Modality Depression Detection Techniques

Automated depression detection has been explored extensively within individual modalities: text, audio, and visual, each capturing distinct aspects of behavior. Studying these modalities independently has not only revealed unique strengths and weaknesses but has also provided essential foundations for designing more sophisticated multimodal systems [15, 21].

Text-Based Approaches

Textual analysis has long been a focal point in the field of computational mental health. Early research examined transcripts from clinical interviews, medical records, and social media platforms [11, 21]. Classical machine learning approaches, including Support Vector Machines (SVM), Logistic Regression, and Random Forests, typically operate on hand-engineered features. These features often include lexical n-grams, sentiment and affect lexicons, or psycholinguistic indicators such as LIWC categories [42, 43].

The introduction of deep learning has expanded the scope of research by capturing sequential dependencies and contextual relationships. Recurrent neural networks (RNNs), Long Short-Term Memory (LSTM) models, and attention-based encoders improve performance by modelling long-range language dynamics [44]. The field advanced further with the advent of pretrained transformer architectures such as BERT [1] and domain-specific variants such as ClinicalBERT [2], which encode nuanced semantics and domain knowledge at scale [45, 30, 46]. More recently, large-scale screening studies on social media (e.g., Twitter, Reddit) have shown the feasibility of population-level monitoring, although they face challenges related to noisy short-form text, demographic bias, and unresolved concerns regarding privacy and consent [47, 48].

Audio-Based Approaches

The acoustic channel provides complementary paralinguistic information that is often disrupted in depression, such as changes in pitch, rhythm, energy, and voice quality [15]. Traditional pipelines focused on handcrafted descriptors such as Mel-Frequency Cepstral Coefficients (MFCCs) or jitter and shimmer, combined with classical classifiers [49].

With deep learning, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been applied to spectrograms and even raw waveforms, improving sensitivity to subtle vocal variations and temporal structures [50, 51]. Benchmark challenges such as AVEC and datasets such as EATD-Corpus have standardized evaluation practices by combining low-level descriptors with statistical functionals and attention-based encoders [52, 30]. Despite these advances, acoustic methods remain vulnerable to environmental noise, microphone variability, and cultural differences in prosody, which complicates generalization in real-world use cases.

Visual-Based Approaches

Visual and facial cues have also proven valuable for depression detection, as symptoms often manifest as reduced expressivity, atypical gaze patterns, and diminished head movement [14, 13, 53]. Early approaches relied heavily on the manual annotation of Facial Action Units (FAUs) and handcrafted computer-vision descriptors.

The field has since moved toward deep feature extraction using CNN architectures (for example, VGG and ResNet), facial landmark tracking tools such as OpenFace, and, more recently, Vision Transformers (ViT), which are capable of capturing fine-grained spatial relationships [54, 36, 55]. Temporal aspects of behavior are often modelled using 3D-CNNs and spatio-temporal networks, which track expression dynamics and head pose trajectories [19]. Research from 2020 onwards has increasingly focused on robustness and explainability in real-world contexts, with studies investigating cultural variability, dataset bias, and systematic failure modes [56, 57]. In this thesis, VGG-16 is adopted as the visual backbone, reflecting its widespread use in earlier affective computing work and its compatibility with simple global average pooling and PCA-based compression, while more recent architectures such as ResNet, MobileNet, or ViT are left as promising alternatives for future exploration.

Limitations of Single-Modality Methods

Despite meaningful advances, systems that rely on a single modality face persistent limitations.

- **Limited coverage** No single modality can capture the full spectrum of depressive symptomatology, which creates blind spots and increases the risk of biased predictions [16].
- **Vulnerability to missing or noisy data** Real-world deployments often face challenges such as poor audio quality, occluded or off-camera faces, and transcription errors, which degrade model reliability.
- **Restricted generalisability** Differences in demographics, language, and recording environments reduce transferability across datasets, thereby limiting robustness [30].
- **Contextual gaps** Some signals, such as sarcasm in text or affective prosody in audio, may not appear in other modalities, leaving context-specific cues undetected in the analysis.

These challenges highlight the need for multimodal fusion and hybrid architectures that combine complementary strengths across modalities. The later sections of this chapter review the progression toward such multimodal systems and the strategies employed to integrate information effectively.

2.3.3 Multimodal Depression Detection

As the limitations of single-modality systems have become evident, multimodal learning has emerged as a central paradigm for automated depression detection. The underlying rationale is that linguistic content, vocal prosody, and nonverbal behavior each capture distinct but complementary facets of affective states. By integrating these heterogeneous signals, multimodal systems can form richer and more reliable representations of depressive symptomatology than

any single stream alone [16, 34, 21]. This section reviews the motivation for multimodal approaches, outlines classical fusion strategies, describes deep learning–based advances, highlights representative applications, and concludes with a discussion of ongoing challenges.

Motivation for Multimodal Approaches

Psychological theory and empirical studies converge on the idea that affective states are expressed through multiple channels simultaneously. In the case of depression, language may reveal increased self-focus or negative valence; acoustic signals may show flattened affect, slowed speech, or abnormal timing; and visual cues such as reduced expressivity, diminished head movement, or atypical gaze reflect changes in social engagement [11, 15, 13]. Because these cues surface under specific contexts or only within particular modalities, systems that rely on a single channel risk overlooking relevant evidence and generating systematic false negatives [25, 19].

From a systems perspective, multimodal modelling also enables graceful degradation; when one channel is missing or corrupted, other modalities can compensate, provided that the fusion mechanism is robust to partial evidence [16]. This redundancy mirrors clinical practice, where clinicians consider verbal, vocal, and behavioral cues together to arrive at a judgment.

Classical Fusion Strategies: Early, Late, and Intermediate

Early computational systems organized multimodal fusion into three main categories. In feature-level (early) fusion, features from different modalities are concatenated before the classification. This approach is conceptually simple and parameter-efficient, but is highly sensitive to scale mismatches, synchronization errors, and missing segments [16].

In early fusion, features extracted from each modality are concatenated into a single joint representation before classification. This can be expressed as

$$z = [z_{\text{text}} \oplus z_{\text{audio}} \oplus z_{\text{visual}}]$$

Here, z_{text}, z_{audio}, and z_{visual} denote modality-specific embeddings, and $\oplus$ represents concatenation. Although this approach is simple and parameter-efficient, it is sensitive to scale mismatches, synchronization errors, and missing data segments [16].

In contrast, decision-level (late) fusion trains unimodal classifiers separately and combines their predictions through averaging, voting, stacking, or learned gating. This strategy improves robustness to missing channels because each stream is processed independently, but it sacrifices fine-grained cross-modal interactions [58, 19].

Intermediate fusion represents a compromise. It employs modality-specific encoders to generate latent representations, which are then integrated through dedicated modules, such as

gating or attention. This design allows both modularity and cross-modal coupling, offering flexibility in handling heterogeneous inputs [16].

The three classical approaches to multimodal fusion (early, late, and intermediate) and more recent deep learning extensions are summarized in Table 2.1.

Table 2.1: Summary of fusion strategies for multi-modal depression detection, highlighting their core design, advantages, and limitations as reported in prior work

Fusion Strategy	**Description**	**Advantages**	**Limitations**
Early Fusion	Feature concatenation before classification	Simple, Parameter-efficient	Sensitive to scale mismatch, missing data [16]
Late Fusion	Combines outputs of unimodal models	Robust to missing channels	Limited cross-modal interactions [58, 19]
Intermediate Fusion	Encoders + shared latent module	Balance of modularity and interaction	Complexity of representation alignment [16]
Deep Learning Fusion	Attention, transformers, memory networks	Strong cross-modal modelling, transferable embeddings	Computationally expensive; domain shift [59, 30]

Deep Learning Approaches to Multimodal Fusion

Modern systems extend classical strategies by leveraging deep-learning encoders and attention-based integration. Convolutional, recurrent, and transformer-based models are commonly used as modality encoders, producing robust embeddings from raw text, acoustic, or visual inputs [15, 14, 1, 2].

Fusion is typically achieved through attention mechanisms, including self-, cross-, and co-attention, which enable models to focus on the most informative segments while down-weighting unreliable signals [59, 60]. Notable architectures include the Memory Fusion Network (MFN), which models cross-view temporal dependencies using memory components [61], and the Multimodal Transformer (MulT) which employs asymmetric cross-modal attention to align modalities at multiple granularities [59]. Subsequent refinements introduced hierarchical attention, structured priors, and domain-informed constraints, reporting significant gains on benchmark datasets such as AVEC and E-DAIC [17, 46, 30].

A recurring theme in these designs is representation alignment. Rather than working with handcrafted features, modern pipelines map heterogeneous modalities to a shared latent

space. Pre-trained language models, such as BERT and ClinicalBERT, provide rich textual embeddings; spectrogram-based CNNs and sequence models yield acoustic representations; and CNN or Vision Transformer (ViT) backbones capture visual and facial features. Fusion modules then operate over these embeddings, improving transferability and reducing the need for manual feature engineering [1, 2, 30, 46]. Transformer-style attention, popularized by Vaswani et al [62], underpins many of

Clinical and Real-World Applications

Empirical studies have consistently demonstrated that multimodal systems outperform unimodal baselines in terms of accuracy and robustness [58, 16]. For instance, Yang et al reported improved performance through audiovisual integration in structured interview settings [19], whereas Shen et al highlighted the benefits of audio–text fusion in the EATD-Corpus [30]. More naturalistic datasets, such as D-Vlog, confirm that multimodal fusion enhances resilience to background noise, variation in head pose, and other real-world artefacts [31].

Beyond raw performance, recent research has emphasized that clinical credibility depends on more than just predictive accuracy. Topics such as explainability, fairness, and deployability are important. Explainability tools help clinicians interpret system decisions [38, 63]; fairness-aware evaluations aim to ensure consistent performance across demographic subgroups [39, 64]; and research on efficient encoders has addressed the operational demands of telehealth and mobile deployments [37, 36].

Limitations and Research Challenges

Despite substantial progress, multimodal depression detection continues to face several unresolved challenges.

- **Missing or noisy modalities** Real-world recordings often contain corrupted or incomplete data. Methods such as training-time modality dropout and robust cross-modal attention have been proposed to mitigate these risks [28, 16].
- **Scalability and efficiency** Deep fusion networks are computationally expensive. Practical deployments in telehealth or mobile contexts require lightweight designs that balance accuracy, latency, and memory constraints [37, 36].
- **Interpretability and calibration** For clinical adoption, models must not only achieve strong performance but also provide interpretable reasoning and calibrated probabilities, with operating points explicitly reported (e.g., sensitivity–specificity trade-offs) [38, 30].
- **Generalisation and domain shift** Performance often degrades when models are applied across languages, cultures or capture settings. Addressing this requires robust

pre-training, domain adaptation techniques, and rigorous cross-corpus evaluation protocols [30, 36].

- **Data availability and annotation** The collection of large-scale multimodal clinical datasets is costly and resource-intensive. Variations in annotation protocols and diagnostic instruments further complicate cross-dataset comparisons and reproducibility [36].

Overall, although multimodal systems represent the state of the art and continue to demonstrate strong empirical performance, their widespread clinical adoption will depend on addressing these challenges alongside sustained attention to fairness, interpretability, and practical deployment.

2.3.4 Attention-Based Fusion Models

Attention has become a cornerstone of modern deep learning and underpins many state-of-the-art multimodal systems in affective computing and mental health [62, 65, 59]. The key idea is to dynamically reweight features, time steps, or entire modalities so that the computational capacity is focused on the most informative signals while unreliable or noisy inputs are down-weighted [66, 34]. For depression detection, attention mechanisms can selectively emphasize diagnostic linguistic markers, prosodic patterns, or facial dynamics at the moments when they are most predictive.

Taxonomy of Attention Mechanisms

Attention mechanisms for multimodal fusion can be broadly categorized into four types: self-attention, cross-modal attention, co-attention, and hierarchical attention. In practice, modern systems often combine multiple forms of data.

- **Self-Attention** Self-attention model dependencies within a single sequence, allowing the system to capture long-range contextual relationships. In depression detection, self-attention has been used within individual modalities, for example, across text tokens, acoustic frames, or facial image sequences to better model the intra-modality structure [1, 67]. Efficiency-oriented variants, such as Sparse Transformers, Linformer, and Performer [68, 69, 70] reduce quadratic complexity, making it feasible to apply attention to long interviews or vlog segments.

  Self-attention allows a model to relate positions within a single sequence, capturing long-range dependency and contextual relationships. Formally, the scaled dot-product attention is defined as

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

Here, Q, K, and V denote the query, key, and value matrices, respectively, and d_k denotes the dimensionality of the keys. This formulation enables the model to compute relevance scores between elements in the sequence and reweight them dynamically, which is particularly valuable for highlighting diagnostic cues in long interviews and transcripts.

- **Cross-Modal Attention** Cross-modal (inter-modal) attention aligns and conditions representations between modalities: one stream (e.g., text) queries another (e.g., visual) to refine its embedding, and vice versa. This mechanism has been particularly effective in affective computing and vision–language models [71, 72, 73]. In depression detection, cross-modal attention enables subtle linguistic cues to be interpreted in the context of concurrent prosodic or facial signals, thereby reducing ambiguity [59, 74, 75].

- **Co-Attention and Hierarchical Attention** Co-attention jointly attends across multiple modalities, learning bidirectional relevance (e.g., how audio influences visual attention and vice versa) [76, 46]. Hierarchical attention applies attention at multiple levels, such as token $\rightarrow$ utterance $\rightarrow$ session, which is especially valuable in depression detection, where local and global contexts may diverge, such as in long clinical interviews or extended vlog narratives [77, 78].

These mechanisms are summarized in Table 2.2, which outlines their purpose, advantages, and limitations as applied in multimodal depression detection.

Table 2.2: Summary of attention mechanisms used in multimodal depression detection, highlighting their purpose, advantages, and limitations as reported in prior work

Attention Type	**Purpose**	**Advantages**	**Limitations**
Self-Attention	Models dependencies within a single modality (e.g., tokens in text, frames in audio/visual)	Captures long-range context; improves intra-modality structure	Computationally heavy for long sequences; mitigated by efficient variants [1, 67, 68, 69]
Cross-Modal Attention	Aligns and conditions one modality on another (e.g., text queries visual/audio)	Disambiguates subtle cues by leveraging complementary streams	Sensitive to misalignment and domain shift [59, 74, 75]
Co-Attention	Jointly attends across two or more modalities bidirectionally	Learns interdependence between modalities; richer integration	Increased model complexity and training cost [76, 46]
Hierarchical Attention	Applies attention at multiple levels (token/segment → utterance → session)	Handles long interactions; captures both local and global context	Requires larger datasets; risk of overfitting on small corpora [77, 78]

Recent Advances in Multimodal Attention Fusion

Transformer-based multimodal architectures have resulted in substantial performance improvements. The Multimodal Transformer (MulT) fuses modalities via asymmetric cross-modal attention pathways [59]; the Memory Fusion Network (MFN) captures cross-view temporal dependencies using gated memory components [61]; and hierarchical transformer variants extend attention to multiple levels for handling long interviews or naturalistic videos [78]. These designs typically rely on pretrained modality-specific encoders, such as language models (e.g., BERT/ClinicalBERT), spectrogram-based CNNs for audio, and CNN or Vision Transformer (ViT) backbones for visual streams, before applying attention layers to integrate them [1, 2, 46, 30]. This reduces dependence on handcrafted features and improves transferability across datasets.

The second theme is **robustness**. Attention modules can adaptively down-weight noisy or missing modalities at runtime (e.g., corrupted audio, off-camera faces) and can be trained with

masking or modality dropout strategies to simulate real-world conditions [60, 28]. Additionally, attention has been linked to interpretability; by highlighting tokens, frames, or modalities that are most influential to a prediction, models become more transparent to end-users [63].

Applications in Affective Computing and Depression Detection

Attention-based fusion has been widely applied across affective computing tasks, including emotion recognition, sentiment analysis, and distress detection, where it consistently improves the benchmark performance [34, 46]. Within depression detection, studies have reported that cross-modal attention not only boosts classification accuracy but also enhances robustness by aligning and weighting linguistic, acoustic, and visual evidence appropriately [30, 79]. In interview-style corpora, audiovisual attention improves tolerance to pose variations, occlusions, and background noise [19]. In more naturalistic "in-the-wild" datasets, such as vlogs, attention helps systems ignore irrelevant segments and focus on expressive moments that are most indicative of depressive symptoms.

Interpretability and Clinical Relevance

An advantage of attention-centric fusion is its amenability to interpretation. Attention maps can be visualized at the token, frame, or modality level, offering clinicians and researchers insight into why a model reaches a given decision [38, 80, 63]. However, recent studies caution that attention weights do not always constitute faithful explanations of internal reasoning [81, 82]. Therefore, the best practice combines attention visualization with complementary explanation methods (e.g., gradient-based attribution) and reports clinically meaningful operating points.

Limitations and Future Directions

Despite their strengths, attention-based models face several challenges. They are often computationally intensive, particularly with long multimodal sequences, motivating ongoing work on efficient transformer variants [68, 83, 69, 70]. Domain shift remains a persistent issue, as attention patterns learned on one dataset may not transfer cleanly to another because of cultural, linguistic, or capture differences [30, 46]. Data scarcity further constrains progress, as attention-heavy models typically require large-scale training corpora, which are limited in clinical mental health contexts [36]. Finally, ensuring accurate explanations and calibrated probability estimates remains critical for trustworthy clinical decision support.

2.3.5 Graph-Based Multimodal Fusion for Depression Detection

Graph Neural Networks (GNNs) have gained traction in affective computing and automated depression detection. Unlike traditional models that treat inputs as independent samples, GNNs

natively represent data as graphs, which are structured collections of nodes and edges. This makes them well-suited for modelling interactions among heterogeneous elements, such as modalities, utterances, or time steps [27, 84]. By explicitly encoding relations, GNNs can capture cross-modal dependencies and temporal structures, both of which are central to mental health analytics, where symptoms and behavioral markers emerge from complex, non-linear interactions.

Relational Learning and Modality Fusion

Graphs can be constructed in several ways for multimodal depression detection. One approach is modality-as-nodes, in which text, audio, and visual embeddings form nodes connected by edges that encode correlations or prior knowledge. Another approach builds segment or utterance graphs, where each node represents a clip or utterance, and the edges capture temporal adjacency or semantic similarity. More complex designs use heterogeneous or bipartite graphs to link different entity types, such as participants and features. Recent studies have demonstrated that these designs outperform early- or late-fusion baselines by enabling explicit relational reasoning [79, 85]. Canonical operators, such as graph convolutions, neighborhood aggregation, and attention-based message passing, enable flexible fusion that is robust to missing features and irregular structures [86, 87, 32].

A standard Graph Convolutional Network (GCN) layer update is expressed as

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)}\right)$$

Here, $H^{(l)}$ represents the node feature matrix at layer l, $\tilde{A}$ is the adjacency matrix with self-loops, $\tilde{D}$ is the degree matrix, $W^{(l)}$ is the trainable weight matrix, and σ is a nonlinear activation function. This operation aggregates information from the neighbors of each node, allowing the model to capture the relational dependencies between modalities or temporal segments. aining robust to missing features and irregular structures.

Temporal Graphs and Dynamic Structure

Depressive symptoms are not static; they evolve over the course of interactions and longitudinal monitoring. Temporal GNN frameworks extend standard GNN by incorporating dynamics, either through continuous-time kernels or snapshot sequences of evolving graphs [88, 89, 90]. For example, in clinical interviews or diary-style recordings, temporal graphs can track how linguistic, acoustic, and facial cues fluctuate over time, capturing symptom trajectories that static models may overlook. Such dynamic relational modeling is particularly relevant in the detection of depression, where momentary cues often combine to form longer-term behavioral patterns.

Interpretability and Clinical Relevance

Graph-based fusion also offers a structured interpretability. Attention coefficients within graph attention networks (GATs) or post hoc methods, such as GNNExplainer, can highlight influential nodes, edges, or features that drive a model's prediction [91, 27]. In multimodal depression detection, this means identifying which utterances, facial expressions, or prosodic patterns most strongly indicate depressive symptoms [92]. This transparency is critical for clinical adoption, as it allows researchers and practitioners to validate model outputs against established behavioral markers.

Current Trends and Open Challenges

Despite the promising progress, several challenges remain. Large graphs with many nodes (e.g., frame-level models) create bottlenecks in scalability. Subgraph sampling and mini-batch training mitigate memory issues but may introduce variance in the estimates [27]. Another issue is edge construction: unusual or arbitrary definitions of relationships can reduce the reliability of the model. Depression datasets are also limited in size and are often labeled with heterogeneous protocols, limiting the training of data-hungry GNNs and complicating cross-corpus comparisons [36]. Domain shift, which is already a challenge in unimodal systems, is amplified in graph settings, where relational structures differ across datasets. Promising directions include self-supervised pre-training for GNNs, contrastive learning, and dynamically learned graph structures that adapt during training [93]. Standardization in reporting, such as clear graph construction rules, edge criteria, and temporal protocols, remains essential for reproducibility.

Graph Transformers and Future Directions

Recently, transformers have been integrated into graph modeling. Graph Transformers combine attention mechanisms with structural encodings, scaling relational learning while retaining the flexibility of transformer-style fusion [94]. Variants incorporate centrality measures, shortest-path distances, and edge attributes into attention computation, thereby capturing richer structural patterns. For depression detection, this points to several opportunities: modeling long interviews with joint graph–token attention, constructing heterogeneous graphs that integrate modalities with clinical metadata, and developing pipelines ready for explanation that combine attention visualization with graph-level attribution [91, 63]. These innovations suggest a path toward clinically credible systems that are both powerful and interpretable.

These design choices are summarized in Table 2.3, which compares different graph construction strategies, their benefits, and the associated limitations in multimodal depression detection.

Table 2.3: Summary of graph-based approaches for multimodal depression detection, highlighting their construction strategy, advantages, and limitations

Graph Type	**Description**	**Advantages**	**Limitations**
Modality Graphs	Each modality (text, audio, visual) represented as a node; edges encode correlations or priors	Explicit cross-modal reasoning; robust to missing modalities	Limited temporal modelling; edge definitions may be arbitrary [79, 85]
Segment/Utterance Graphs	Nodes represent utterances or clips; edges encode temporal adjacency or semantic similarity	Captures local temporal context; interpretable at segment level	Large graphs for long sessions; edge noise can degrade performance [86, 87]
Heterogeneous Graphs	Different entity types (modalities, speakers, metadata) form nodes; bipartite or typed edges link them	Flexible; models multi-entity interactions; integrates metadata	Complex design; requires careful edge schema [32, 27]
Temporal/Dynamic Graphs	Graph structure evolves over time via snapshots or continuous-time kernels	Models progression of depressive cues; captures symptom trajectories	High computational cost; requires sequential data [88, 89, 90]
Graph Transformers	Combines attention with graph encodings (centrality, path, edge features)	Scales relational modelling; unifies graphs with transformer flexibility	Data hungry; relatively unexplored for clinical use [94]

Most of the approaches reviewed in this chapter have been evaluated on a small number of benchmark datasets, such as DAIC-WOZ, AVEC, EATD-Corpus, and D-Vlog. These corpora differ in scale, capture environment, and labelling protocols, which have implications for the generalizability and comparability of the results. A detailed description of the datasets used in this study is provided in Chapter 4.

2.4 Summary

This chapter reviews the landscape of depression detection, outlining its multimodal nature, practical classification schemes, and key challenges involved. This highlights the clinical importance of reliable screening and traces the methodological evolution from manual and

instrument-based assessments to single-modality computational methods, and finally to multimodal fusion approaches incorporating attention mechanisms and graph-structured models.

Recurrent themes across the literature include the need for robustness to noisy or missing inputs, alignment across modalities, generalization across domains, and transparency with calibrated outputs for clinical use. These observations inform the dataset choices presented in Chapter 4 and motivate the modelling architecture described in Chapter 3.

Chapter 3

MLGA - A Modality-Level Graph Attention Architecture for Multimodal Depression Detection

---

This section introduces the proposed Modality-Level Graph Attention (MLGA) framework for multimodal depression detection. We begin with a high-level overview of the pipeline and its design rationale, followed by concise descriptions of the three input modalities, text, visual, and facial, used throughout this work. All remaining architectural elements (regularization, shared-space projection, graph-attention fusion, temporal modelling, and the classifier) are briefly summarized in prose here to preserve continuity; full details and ablations are provided in subsequent chapters.

3.1 Overview of MLGA

The proposed Modality-Level Graph Attention (MLGA) architecture is a hybrid model that integrates specialized modality encoders with graph-based fusion and lightweight temporal reasoning. In broad terms, the system comprises two stages: (i) modality-specific feature extraction, which converts raw artifacts of an interview (transcripts, video frames, and frame-wise facial descriptors) into compact segment-level embeddings; and (ii) modality-level reasoning, which aligns these heterogeneous vectors in a shared space, performs cross-modal aggregation with graph attention, and models within-session dynamics before producing a calibrated depression probability. Figures 3.4 and 3.5 illustrate the high-level pipeline and block-level data flow, respectively.

Given an interview, the audio–video stream is segmented, and each segment is processed through three parallel channels. The textual stream is encoded with ClinicalBERT to obtain 768-dimensional contextual embeddings that capture the semantic and pragmatic cues associated with depressive expression [1, 2]. In parallel, the visual stream is passed through a VGG-16 backbone pretrained on ImageNet to produce 512-d appearance descriptors, which may be PCA-compressed to 128-d for efficiency [3, 4]. The facial stream is represented as a numerical time-series derived with OpenFace (Action Units, landmarks, gaze, head pose); per-segment

summary statistics yield a compact 16-d vector reflecting FACS-based behaviors [54, 95]. These three embeddings, text, visual, and facial, constitute the raw modality features for that segment (detailed in the subsequent subsections).

Before cross-modal integration, each modality vector is L2-normalised, and robustness mechanisms are applied during training: zero-mean Gaussian noise encourages tolerance to sensor and preprocessing variability [96], while modality dropout randomly masks the entire modality vectors at the segment level to promote graceful handling of missing or degraded channels [28]. The normalized (and, at training time, perturbed) features are then mapped by a trainable linear layer with ReLU into a common 512-dimensional latent space and renormalised. This alignment produces comparable node features across modalities, preserving the informative structure from high-capacity encoders (ClinicalBERT/VGG) without overcompression and preparing the representations for attention-based fusion.

Cross-modal reasoning is realized through a fully connected three-node modality graph whose nodes correspond to $\{$Text, Visual, Facial$\}$. A two-layer Graph Attention Network (GAT) computes data-dependent attention coefficients that regulate information flow along edges (e.g., text $\rightarrow$ face, visual $\rightarrow$ text), thereby focusing aggregation on the most informative inter-modality relations [32]. Attention coefficients provide concise modality-level attributions that aid interpretability. The resulting fused sequence of segment representations is then passed to a lightweight, single-layer GRU that encodes temporal dependencies across the interview [97]. Variable-length sessions are supported via padding masks, and dropout is applied to the inputs during training. The final hidden state is mapped through a sigmoid unit to produce the depression probability, optimized with (optionally class-weighted) binary cross-entropy. For reporting, a validation-selected threshold (F1-oriented) converts probabilities to labels, and the results are accompanied by PR/ROC curves and confusion matrices to characterize precision–recall trade-offs under class imbalance [98]. At inference, the inputs are unperturbed, and any absent modality is simply masked, ensuring robust behavior in telehealth and in-the-wild conditions.

3.1.1 Textual Modality

The textual channel is processed using ClinicalBERT, a domain-adapted version of the BERT architecture pretrained on large collections of clinical notes [2]. This adaptation improves its ability to represent medical and psychological terminology beyond what is captured by the standard BERT model [1]. The base configuration consists of 12 transformer layers, 12 attention heads, and a hidden size of 768, supported by WordPiece tokenization and positional/segment embeddings (Figure 3.1).

In our framework, the participant transcripts were tokenized, normalized, and padded to a fixed sequence length. Attention masks were constructed to distinguish between valid tokens

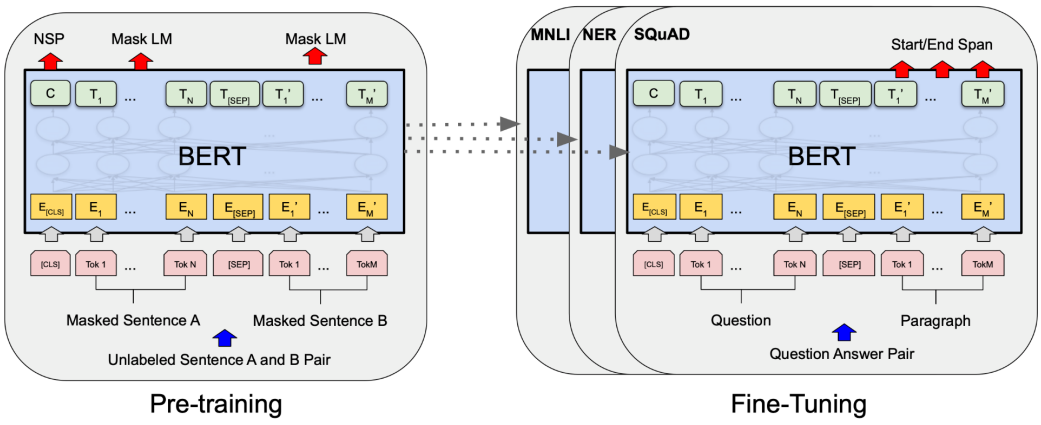

Figure 3.1: BERT-base architecture used as the backbone for the text encoder (12 layers, 12 heads, hidden size 768), including WordPiece tokenization and token/segment/position embeddings. Diagram created by the author; architecture per [1]

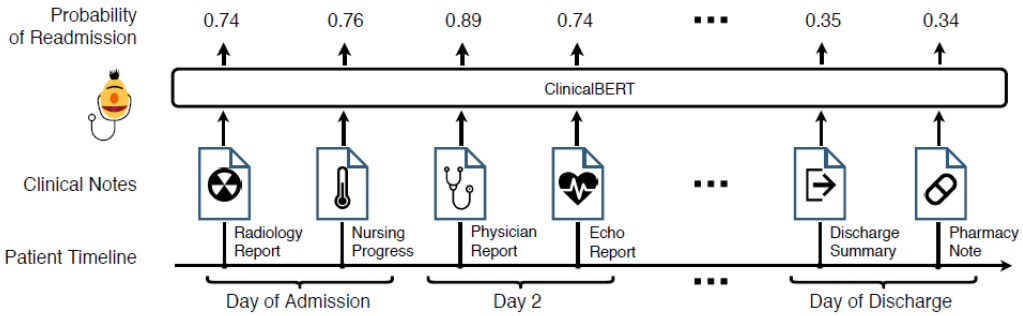

Figure 3.2: ClinicalBERT text encoder used in our model. The network shares the BERT-base configuration but is domain-adapted via pretraining on clinical notes. We used the [CLS] embedding for the binary depression classification. Diagram created by the author; domain adaptation per [2]

and padding positions. The [CLS] embedding produced by ClinicalBERT was then extracted for each segment and fine-tuned with a task-specific classification head (Figure 3.2). These embeddings capture semantic and syntactic cues in language, such as negative affect, cognitive distortions, and repetitive self-focus, which have been strongly linked to depressive symptomatology [11, 12]. The resulting representations are subsequently passed into the fusion and temporal modules (Sections 3.4–3.5.2).

3.1.2 Visual Modality

Visual features are derived from raw video frames using a VGG-16 convolutional neural network backbone pretrained on the ImageNet dataset [3, 4]. Following the transfer learning practice, only the convolutional stack up to the pool5 layer is retained, discarding the original classifier.

Each frame produced a tensor of size 7×7×512 at the pool5 stage, which was then reduced through global average pooling (GAP) into a 512-dimensional descriptor.

To standardize storage and reduce redundancy prior to downstream modeling, these descriptors are optionally compressed to 128 dimensions using Principal Component Analysis (PCA) [99]. In our pipeline, the visual input handed to preprocessing is a 128-dimensional vector per segment. The complete extraction pathway is illustrated in Figure 3.3.

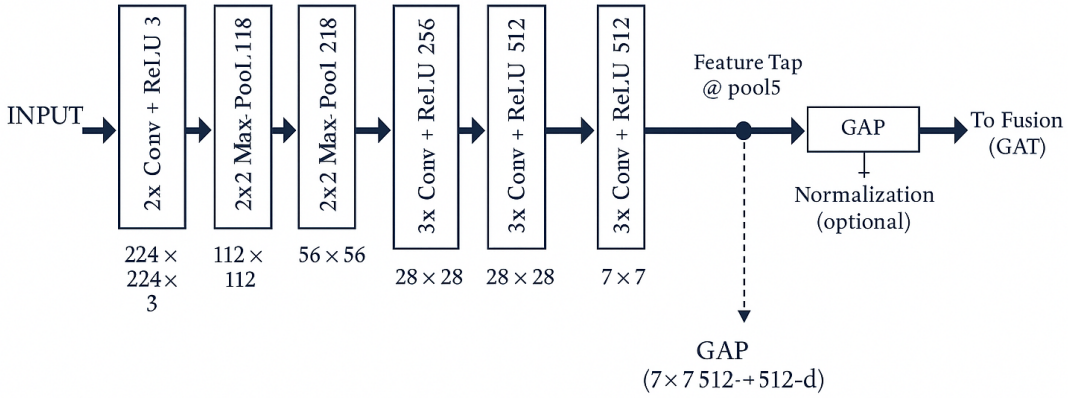

Figure 3.3: VGG-16 visual feature extractor. The convolutional stack is retained up to pool5; the pool5 tensor (7×7×512) is aggregated via GAP to a 512-dimensional vector, optionally PCA-compressed to 128 dimensions before preprocessing and projection. Architecture per [3]; pretraining on ImageNet per [4].

From a design perspective, VGG-16 was selected as the visual backbone for three reasons. First, it produces dense convolutional feature maps with a simple and well-understood structure, which are amenable to global average pooling and subsequent PCA compression into low-dimensional descriptors, aligning with the goal of keeping the downstream fusion stack compact. Second, VGG-16 is widely used in earlier affective computing and medical imaging work, which facilitates comparison with existing baselines and reuse of established implementation practices. Third, freezing a pretrained VGG-16 and operating only on its intermediate features reduces the engineering effort and computational cost relative to fully retraining more recent architectures. Nonetheless, VGG-16 is an older and comparatively heavy backbone, and the use of more modern, parameter-efficient visual encoders (for example, lightweight ResNet or MobileNet variants) is a natural direction for future work.

3.1.3 Facial Modality

The facial modality is represented as numerical time-series features rather than raw image data. Frame-level facial descriptors, including Action Units (AUs), 2D/3D landmarks, gaze direction, and head pose, were extracted using the OpenFace toolkit [54, 95]. Frames with low confidence were discarded to ensure reliability, and each feature channel was z-score normalized across time.

Let $X \in \mathbb{R}^{T \times K}$ denote the AU feature matrix for a segment, where T is the number of frames and K is the number of selected AUs. For each channel, a temporal summary statistic (mean activation across frames) was computed as follows:

$$z_k = \frac{1}{T} \sum_{t=1}^{T} X_{t,k}, \quad k = 1, \ldots, 16.$$

This produces a compact 16-dimensional vector per segment, reflecting clinically relevant facial behaviors defined by the Facial Action Coding System (FACS) [100]. Typical indicators include inner brow raises, brow lowering, eyelid tightening, and lip corner pulling. The 16-dimensional vector was then forwarded to the preprocessing stage (Section 3.3) and subsequently projected to the shared 512-dimensional space (Section 3.2).

As shown in figure 3.4 depicts the schematic overview of the framework, and figure 3.5 visualizes the block-level architecture showing the modality-specific inputs, normalization, regularization, GAT-based fusion, GRU modeling, and final classification.

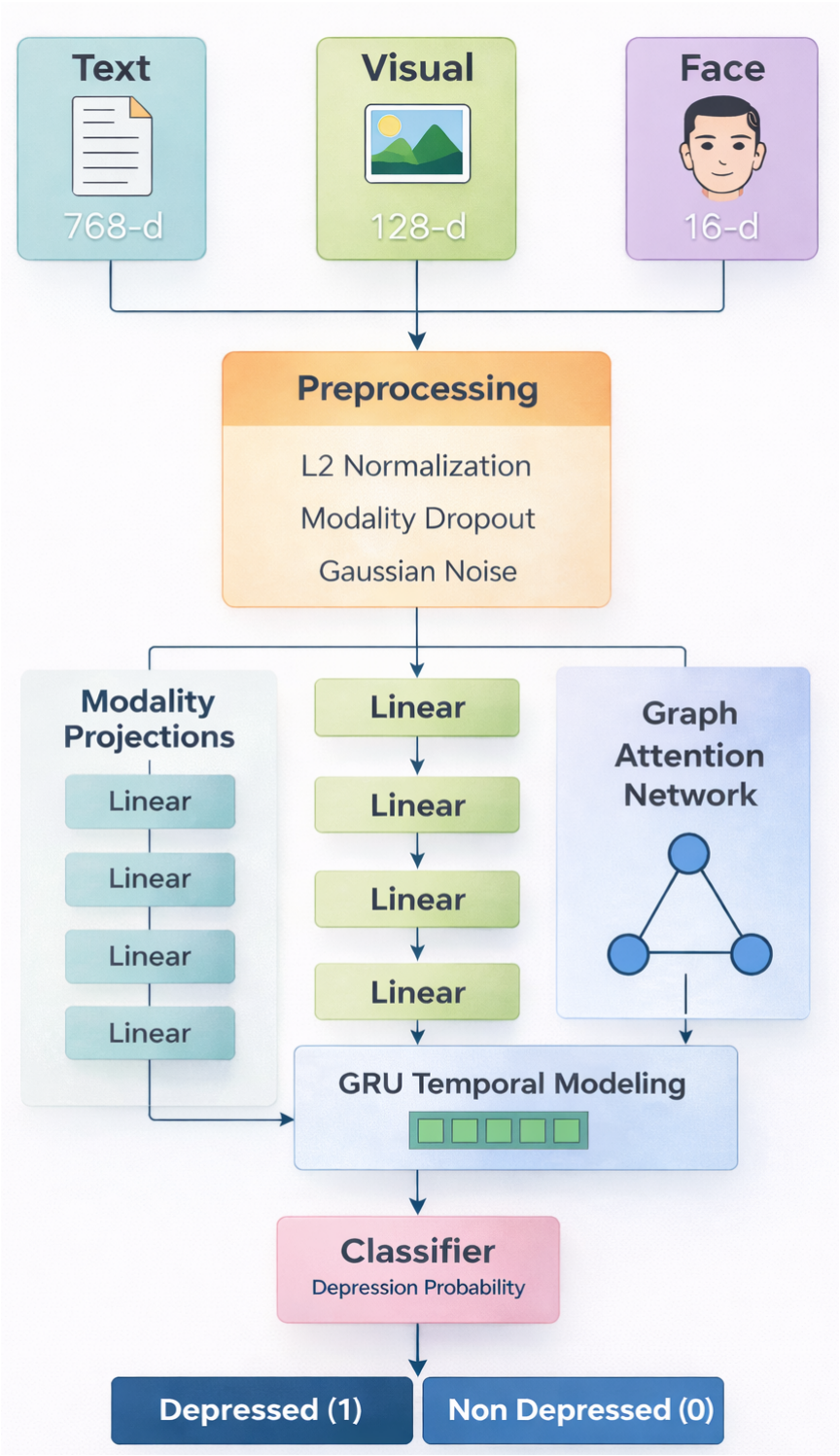

Figure 3.4: High-level pipeline of the proposed multimodal architecture integrating Clinical-BERT, VGG-16, OpenFace, GAT, and GRU modules

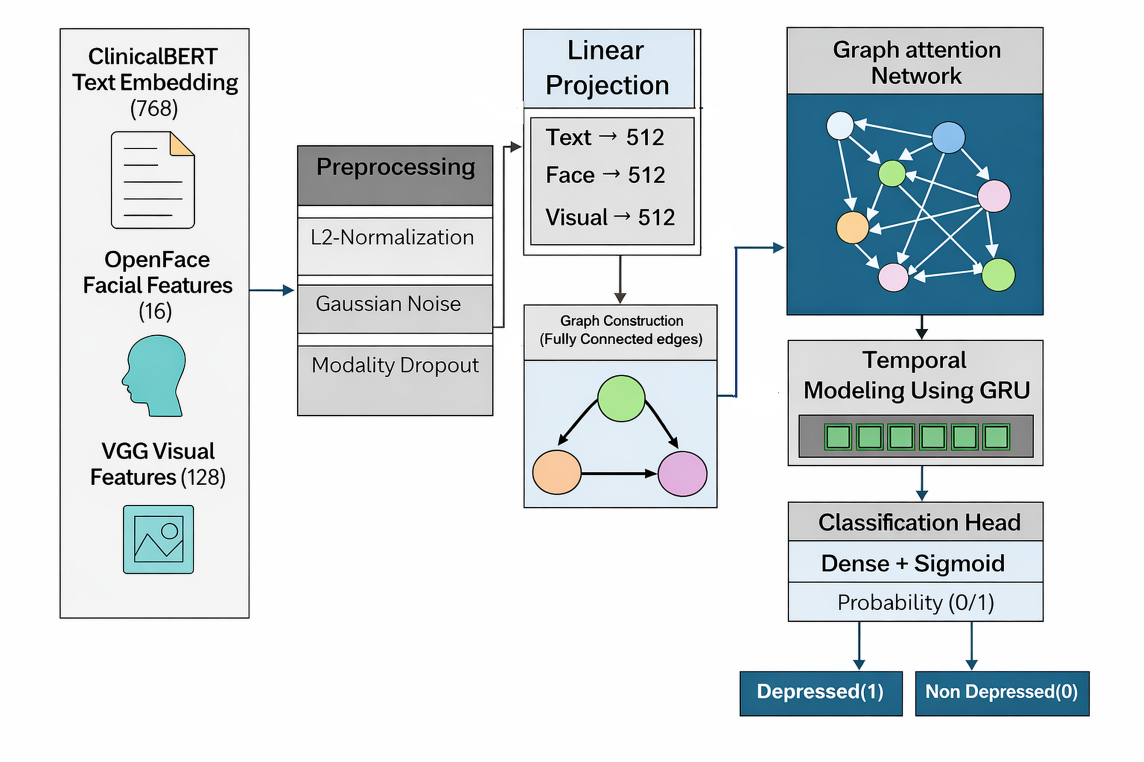

Figure 3.5: Detailed block-wise architecture showing modality extraction, projection, dropout regularization, graph attention fusion, temporal modeling via GRU, and final classification

Component Summary

1. *Feature Extraction* Text (ClinicalBERT), visual (VGG-16+PCA), and facial (OpenFace) embeddings are extracted.

2. *Projection into Shared Latent Space* All modality features are normalized and projected into a unified latent representation.

3. *Regularization* Gaussian noise and modality dropout improve generalization and robustness.

4. *Graph Attention Fusion* Cross-modal relations are modeled using GAT layers with learnable attention scores.

5. *Temporal Modeling* GRU layers encode the sequential dynamics of fused multimodal features.

6. *Classification* A fully connected classifier with sigmoid activation maps the representation to depression probability.

This modular pipeline enables end-to-end training, interpretable attention-driven fusion, and flexible integration of additional modalities into the model. The use of standardized feature

extractors (ClinicalBERT, VGG-16, OpenFace) with advanced deep learning modules aligns with current best practices in affective computing and health informatics [101, 17, 30].

3.2 Modality-Specific Embedding Representation and Projection

For multimodal integration to be effective, heterogeneous features from the text, visual, and facial streams must be projected into a shared latent space. In this study, all modalities were mapped into a 512-dimensional common embedding space after preprocessing (L2-normalization, Gaussian noise, and modality dropout).

Feature sources and dimensional characteristics:

- **ClinicalBERT (text)**: 768-dimensional contextual embeddings for each segment [2].
- **VGG (visual)**: 512-dimensional descriptors from convolutional activations, optionally reduced to 128 dimensions through PCA; the 128-d vector is the visual input to preprocessing.
- **OpenFace (facial AUs)**: compact 16-dimensional descriptors derived from a selected subset of AUs [54].

Projection layers

Each modality m (with post-preprocessing dimensionality $d_m \in \{768, 128, 16\}$) is mapped to the shared 512-dimensional space via a modality-specific linear layer followed by a nonlinearity:

$$\hat{x}_m = \sigma(W_m \bar{x}_m + b_m), \qquad W_m \in \mathbb{R}^{512 \times d_m},\ b_m \in \mathbb{R}^{512},$$

where $\bar{x}_m$ denotes the preprocessed (L2-normalized, noise-perturbed, and possibly dropped) modality vector, and σ is the ReLU. We then applied post-projection normalization (LayerNorm) to stabilize the scale across modalities:

$$\tilde{x}_m = \text{LayerNorm}(\hat{x}_m), \qquad \tilde{x}_m \in \mathbb{R}^{512},$$

yielding the node features used for graph construction (Section 3.4). Parameters $\{W_m, b_m\}$ are not shared across modalities, allowing for modality-aware capacity while aligning them in a common space.

Rationale for 512 dimensions

A 512-dimensional latent space preserves rich information from high-capacity ClinicalBERT (768-d) and VGG (512-d/128-d PCA) while avoiding overcompression. It provides a stable and expressive basis for attention-based fusion (GAT) and temporal modeling (GRU). Robustness is enforced at the modality level (pre-projection) by Gaussian noise and modality dropout.

Dimensional summary

Table 3.1: Input and projected dimensions for each modality. Visual features may be PCA-compressed to 128-d prior to preprocessing; all modalities are projected to 512-d for fusion

Modality	Input Dimension to Preprocessing	Projected Dimension
ClinicalBERT Text	768	512
OpenFace Facial Features	16	512
VGG Visual Features (PCA)	128	512

The projected 512-dimensional embeddings form the node features in the modality-level graph (Section 3.4). The GAT performs cross-modal fusion and outputs a 128-dimensional fused sequence that is modeled by the GRU (Section 3.5.2).

3.3 Regularization via Gaussian Noise and Modality Dropout

Deep learning models are often prone to overfitting, particularly when trained on heterogeneous data from multiple sources. In the context of multimodal depression detection, this challenge is amplified because each modality (text, visual, and facial) differs in terms of dimensionality, scale, and noise characteristics. To ensure that the learned representations are robust and generalizable across participants and recording conditions, the proposed architecture incorporates three regularization steps before projection into the shared 512-dimensional space: L2-normalization, Gaussian noise injection, and modality dropout. Together, these steps stabilize the training, reduce the sensitivity to spurious variations, and prepare the embeddings for effective cross-modal fusion.

3.3.1 L2-Normalization

For each modality vector x_m, we first applied L2-normalization:

$$\hat{x}_m = \frac{x_m}{\|x_m\|_2 + \epsilon},$$

where a small constant ϵ prevents the division by zero. This step ensured that embeddings from different modalities (text: 768-d, visual: 128-d, and facial: 16-d) were mapped onto comparable scales. Without this step, one modality with a higher raw magnitude (e.g., text embeddings) could dominate others during training, leading to biased fusion. Normalization also improves the stability of the optimization because all features contribute within a consistent range.

3.3.2 Gaussian Noise Injection

To further enhance robustness, Gaussian noise was added to the normalized embeddings.

$$\bar{x}_m \;=\; \hat{x}_m + \varepsilon_m, \qquad \varepsilon_m \sim \mathcal{N}\Big(0,\, \sigma^2 I_{d_m}\Big),$$

where $d_m \in \{768, 128, 16\}$ is the dimensionality of the modality, and σ=0.1 unless specified otherwise. Intuitively, this means that each embedding is slightly perturbed during training. By learning under noisy conditions, the model is discouraged from memorizing exact feature patterns and instead learns more generalizable representations of the data. This is especially important in depression detection, where input variability may arise from differences in the language style, lighting, facial pose, or background noise. Gaussian noise acts as a data-level augmentation that improves the resilience to such perturbations [96, 102].

3.3.3 Modality Dropout

The final step is modality dropout, which operates at the embedding level.

$$\tilde{x}_m \;=\; m_m \cdot \bar{x}_m, \qquad m_m \sim \text{Bernoulli}(p),$$

with a retention probability p=0.8. In practice, this means that during training, each modality (text, visual, or facial) has a 20% chance of being entirely dropped. By randomly masking complete modalities, the network is forced to learn complementary relationships across streams rather than relying on a single dominant channel. For example, if textual information is missing or corrupted in a telehealth setting, the model should still be able to infer depression symptoms from facial or visual cues. This strategy directly simulates real-world conditions in which not all modalities are consistently available [28].

Implementation and Hyperparameters

The preprocessing and regularization strategy combined L2-normalization, Gaussian noise injection, and modality dropout prior to projection into the shared latent space. These steps were consistently applied across all modalities to stabilize the training and enforce robustness. Table 3.2 summarizes the hyperparameter settings, and Figure 3.6 visually illustrates the process applied to each modality. Together, these operations ensured that the embeddings (text: 768-d, visual: 128-d PCA, facial: 16-d) were normalized, noise-perturbed, and randomly masked with retention probability p=0.8, preparing them for projection into the common 512-d space (Section 3.2). This procedure improved generalization, reduced overfitting, and enhanced the resilience to missing or degraded inputs.

Table 3.2: Preprocessing and regularization hyperparameters used prior to projection

Technique	**Setting**
L2-Normalization	Applied per modality vector
Gaussian Noise Std. (σ)	0.1
Modality Dropout Retention (p)	0.8

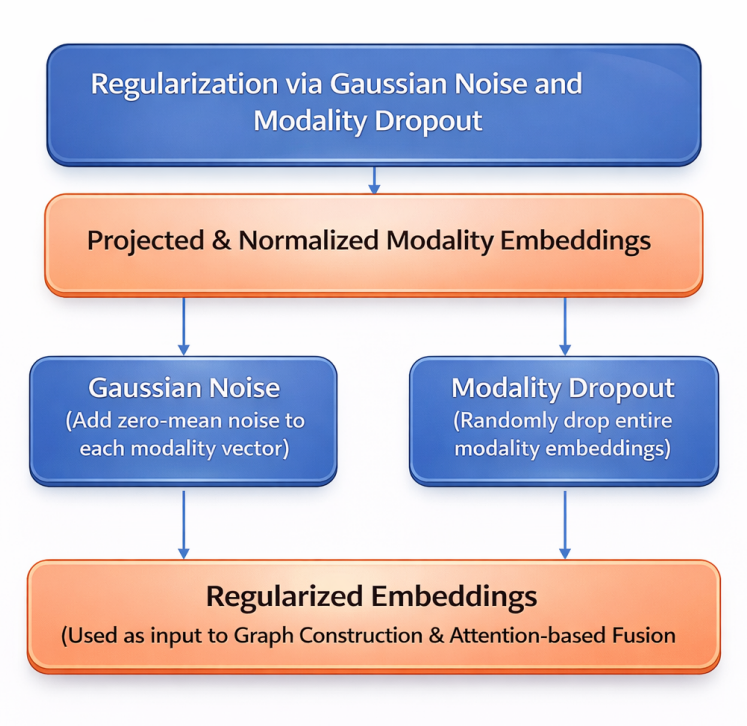

Figure 3.6: Pre-projection regularization process. Each modality embedding (text: 768-d, visual: 128-d, facial: 16-d) was first L2-normalized, then perturbed with Gaussian noise (σ=0.1), and subjected to modality dropout (p=0.8) before projection into the shared 512-d latent space. Diagram created by the author

3.4 Graph Construction with Graph Attention Networks

After preprocessing (Section 3.3) and projection to the shared latent space (Section 3.2), each modality embedding is represented as a node in a fully connected graph. In this formulation, the set of nodes is $V = \{\text{text}, \text{visual}, \text{facial}\}$ with $|V|$=3, and each node is initialized with a 512-dimensional feature vector derived from the corresponding modality-embedding. Directed edges connect every pair of nodes, allowing the network to explicitly model the cross-modal interactions. A Graph Attention Network (GAT) [32] is then applied to adaptively weight information exchanged along these edges, enabling the model to emphasize the most informative cross-modal relationships for depression detection [103].

The graph construction process is formalized in Algorithm 1, where the projected modality embeddings are converted into nodes and connected via directed edges to form a fully connected graph. This structure is subsequently processed by the GAT layers for an adaptive cross-modal fusion.

Algorithm 1 Graph Construction from Projected Modality Embeddings

Function CreateGraph(projected embeddings)
Initialize empty graph G
for each modality embedding x_i in projected embeddings **do**
    node ← CreateNode(x_i)
    AddNode(G, node)
end for
for each pair of nodes (v_i, v_j) where $i \neq j$ **do**
    AddDirectedEdge(G, v_i, v_j)
end for
return graph G with fully connected modality nodes

Mathematical Formulation of GAT

Let $\{h_i\}_{i \in V}$ denote the set of modality embeddings, with each $h_i \in \mathbb{R}^{\mathbf{512}}$. Each embedding is first linearly transformed as follows:

$$h'_i = W h_i, \qquad W \in \mathbb{R}^{d' \times \mathbf{512}},$$

where d' denotes the head dimension. For an edge from node j to node i, a raw attention score is computed as

$$e_{ij} = \text{LeakyReLU}\Big(\mathbf{a}^\top \Big[\, h'_i \,\|\, h'_j \,\Big]\Big),$$

where $\mathbf{a} \in \mathbb{R}^{2d'}$ is a learnable attention vector, and $\|$ denotes concatenation. These raw scores are normalized across the neighborhood $\mathcal{N}(i)$ using the softmax function as follows:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}(i)} \exp(e_{ik})}.$$

The updated node representation is then given by the attention-weighted sum of its neighbors as follows:

$$h''_i = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}\, h'_j \right),$$

with σ a nonlinearity (ELU or ReLU). In the multi-head setting with K heads, independent sets of parameters $\{W^{(k)}, \mathbf{a}^{(k)}\}_{k=1}^{K}$ are used, and head outputs are concatenated or averaged, as in [32]. Intuitively, these equations assign a learnable importance score to each edge, determining the extent to which one modality should influence another at a given time step.

After message passing through the stacked GAT layers, the modality-specific embeddings (each 512-d) are aggregated and compressed into a single 128-dimensional fused vector per time step. This dimensionality reduction distills the most salient cross-modal cues while maintaining a compact representation for temporal modeling. The resulting sequence $\{u_t\}_{t=1}^{T}$ with $u_t \in \mathbb{R}^{128}$ is forwarded to the GRU module (Section 3.5.2) for the dynamic analysis of interview segments.

The inputs to the GAT were the 512-d modality embeddings after preprocessing and projection. If the modality dropout masked a modality during preprocessing, its projected vector was set to zero, and attention was redistributed across the remaining neighbors. The standard GAT dropout, which is applied to node features and attention coefficients, operates independently of the modality dropout and provides additional regularization inside the GAT layers.

As shown in figure 3.7 illustrates the conceptual mechanism of GAT-based fusion: edges between nodes are weighted by attention scores, highlighting the modalities that contribute most strongly to one another. For example, text features may provide a more informative context for visual cues in one segment, whereas facial features may dominate in another. These dynamic weights render fusion interpretable and adaptive.

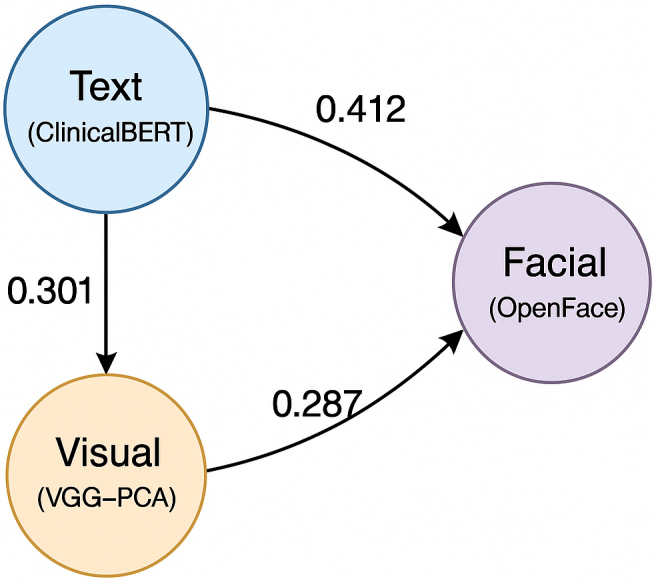

Figure 3.7: Attention-based modality fusion. Edges between modality nodes are adaptively weighted by the GAT, emphasizing cross-modal relationships relevant for depression detection

As shown in figure 3.8 presents the block-level workflow of the GAT module. The "block" here refers to the internal operations inside the GAT: each node begins as a 512-d modality embedding, a fully connected graph is constructed, stacked GAT layers compute attention and update node embeddings, and the outputs are compressed into a fused 128-d vector. Therefore, this block bridges modality-specific representations and the temporal GRU encoder, providing a unified multimodal representation at each time step.

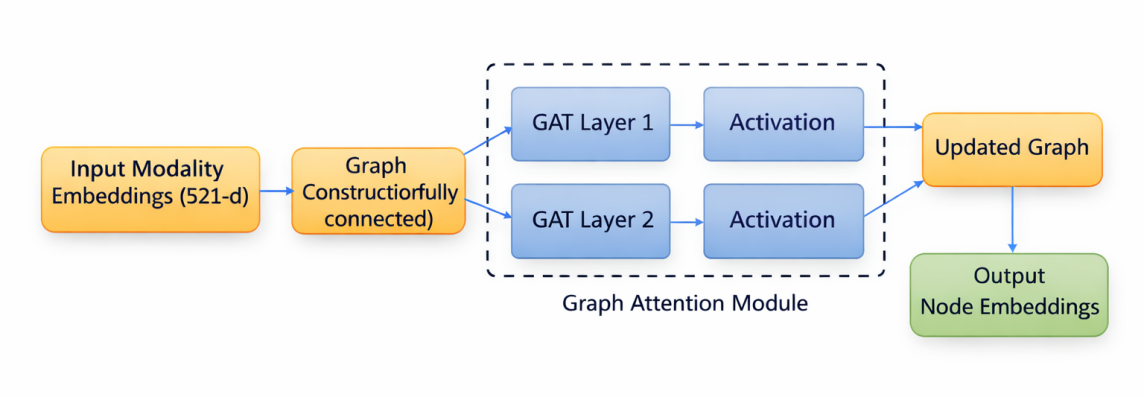

Figure 3.8: Block diagram of the GAT module. Each node starts as a 512-d modality embedding; stacked GAT layers with non-linear activations and in-layer dropout produce a fused 128-d representation per timestep

3.5 Fusion Mechanism and Classifier Head

After the modality-specific embeddings are extracted and projected into a shared latent space, the Graph Attention Network (GAT) layers refine them by explicitly modeling cross-modal relationships. These updated modality embeddings are then combined, compressed, and passed through the temporal modeling and classification modules. Collectively, these stages form the decision-making core of the framework, converting raw multimodal descriptors into participant-level probabilities of depression. figure 3.9 provides a schematic overview of the fusion-to-classification pipeline.

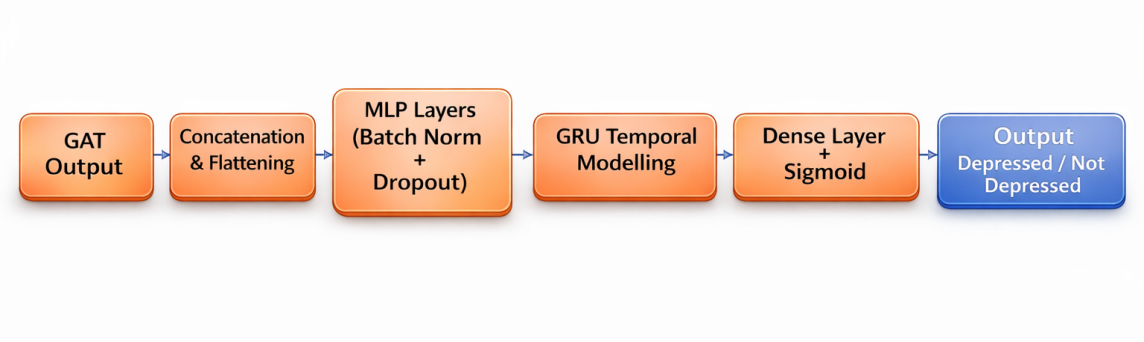

Figure 3.9: Fusion and classification pipeline. Node embeddings from the GAT are concatenated, refined using MLP layers with Batch Normalization and Dropout, modeled temporally via a GRU, and classified through a dense layer with sigmoid activation

3.5.1 Fusion Process

The outputs of the final GAT layer are 512-dimensional modality embeddings, one per modality node, which are updated through attention-weighted message passing. These embeddings, which capture how each modality (text, visual, facial) influences and complements the others, are concatenated to create a unified representation. This concatenated vector was then compressed into a compact 128-dimensional fused embedding for each time step.

The motivation for this compression is twofold: (i) it distills the most salient cross-modal information while filtering redundancy, and (ii) it reduces dimensionality, lowering the computational cost for downstream processing. Importantly, the fused embedding encodes subtle interactions, such as how linguistic cues may align with facial expressivity, providing a richer signal than any modality in isolation.

This 128-d fused embedding is then passed through a two-layer Multilayer Perceptron (MLP), which further captures higher-order nonlinear interactions. To ensure stability and generalization, two forms of regularization were applied to each layer:

- **Batch Normalization** standardizes intermediate activations, reducing internal covariate shift and allowing faster, more stable convergence during training [104].
- **Dropout** randomly deactivates neurons during training to discourage co-adaptation of features, ensuring robustness when different modality signals dominate in different participants [105].

This combination ensures that the fused representation is not only compact and informative but also resilient to overfitting and spurious correlations.

3.5.2 Temporal Modeling

While the fused embeddings summarize multimodal evidence at each time step (segment), depression cues often evolve over the course of an interview; patients may begin guarded, reveal more over time, or fluctuate in affect. To capture such dynamics, we used a Gated Recurrent Unit (GRU), a recurrent neural network variant optimized for long sequences.

At each step t, the GRU updates its hidden state h_t based on the current fused embedding $x_t \in \mathbb{R}^{128}$ and the previous hidden state h_{t-1}. Its gating mechanisms regulate the amount of past information to retain and the amount to incorporate from the current input, mitigating the vanishing gradient issues that affect traditional RNNs. Formally:

$$h_t = \mathrm{GRU}(x_t, h_{t-1}), \qquad h_0 = \mathbf{0}.$$

Variable-length sessions are supported by padding sequences and the application of binary masks so that padded steps do not contribute to the loss. Dropout is also applied to GRU inputs

during training, complementing the modality-level robustness strategies (Gaussian noise and modality dropout in Section 3.3).

The GRU produces a final hidden state $h_T \in \mathbb{R}^H$, which condenses the sequential trajectory of multimodal signals into a session-level representation, effectively summarizing the participant's entire interaction.

3.5.3 Classification Head

The session representation h_T is then mapped to the probability of depression through a dense layer with sigmoid activation:

$$\hat{y} = \sigma(w^\top h_T + b), \qquad w \in \mathbb{R}^H,\ b \in \mathbb{R}.$$

This produces $\hat{y} \in [0, 1]$, which is interpreted as the likelihood that the participant is depressed. Training uses binary cross-entropy loss, optionally weighted to correct for class imbalance (depressed vs. non-depressed participants).

At inference, the continuous score $\hat{y}$ is thresholded to generate a binary label. This allows the system to operate as either a strict classifier (binary outputs) or a calibrated risk scorer (continuous probabilities), depending on the application.

3.5.4 Threshold Optimization

Rather than adopting the default 0.5 cutoff for $\hat{y}$, the threshold was tuned on the validation set to maximize the F1-score [98]. This balances the sensitivity (recall) and specificity (precision), ensuring that the model performs well under the clinical reality of imbalanced datasets. Optimizing the threshold is critical.

- A threshold that is too low may inflate false positives, creating unnecessary concern or overdiagnosis.
- A threshold that is too high may miss true cases, delaying needed intervention.

By explicitly tuning the threshold, the framework aligns with the dual imperative of clinical safety (catching as many true cases as possible) and clinical trust (avoiding spurious alarms).

3.6 Summary

This chapter introduces the proposed multimodal framework for automated depression detection, outlining its architectural design, regularization strategies, and fusion mechanisms. The framework integrates textual (ClinicalBERT 768-d), visual (VGG-16 with optional PCA to 128-d), and facial (OpenFace 16-d) data streams.

We first applied L2-normalization, Gaussian noise injection, and modality dropout before projection to encourage modality-level robustness. The preprocessed vectors were then projected onto a unified 512-dimensional latent space, forming the node features of a fully connected modality graph. Graph Attention Networks (GAT) dynamically model cross-modal dependencies and produce a fused 128-dimensional sequence that a GRU encodes over time. A sigmoid-activated dense layer outputs the depression probability, and the decision threshold is optimized on the validation set to maximize the F1 score.

This design mirrors the architectural schematic and provides a robust, interpretable pipeline in which the fusion, graph, temporal, and classification components are trained jointly on top of fixed pretrained embeddings. Empirical validation is presented in Chapter 6.

Chapter 4

Datasets and Preprocessing

4.1 Datasets Overview and Collection

This study draws on three widely used corpora for automated depression detection: the E-DAIC-WOZ, EATD-Corpus, and D-Vlog datasets. Taken together, they span structured clinical interviews and naturalistic video content, which enables evaluation under both controlled and wild conditions.

4.1.1 E-DAIC-WOZ

The Extended Distress Analysis Interview Corpus Wizard-of-Oz (E-DAIC-WOZ), released as part of the AVEC 2019 challenge [106], is an extended version of the original Distress Analysis Interview Corpus (DAIC) introduced by Gratch et al. [107]. The dataset consists of semi-structured clinical interviews conducted by a virtual agent, "Ellie," designed to elicit behaviorally and clinically relevant cues associated with psychological distress and depression.

- **Participants** 275 individuals aged 16–60, with mixed gender representation and varying degrees of depressive severity.
- **Modalities**
  - **Textual** Sentence-level transcripts of participant speech during interviews.
  - **Visual** Video frames sampled at 1 fps, later used for VGG-based embedding extraction.
  - **Facial** Behavioural features extracted via OpenFace [54], including facial Action Units (AUs), gaze direction, and head pose, captured at 30 fps.
- **Annotations** Depression severity was annotated using the Patient Health Questionnaire (PHQ-8) [8], with scores ranging from 0 to 24. Following clinical guidelines, binary diagnostic labels were assigned as follows:

$$\text{Label} = \begin{cases} 1, & \text{if PHQ-8} \geq 10 \text{ (depressed)} \\ 0, & \text{otherwise (non-depressed).} \end{cases}$$

  A PHQ-8 score of 10 or higher is clinically indicative of moderate-to-severe depression.

The E-DAIC-WOZ has since become a benchmark resource in affective computing and multimodal learning, offering structured, clinically relevant interviews that are particularly well-suited for evaluating models integrating textual, visual, and behavioural signals [101, 34].

4.1.2 EATD-Corpus

The Emotional Audio-Textual Depression Corpus (EATD-Corpus) [30] is a Mandarin dataset designed to study the linguistic and prosodic correlates of depression.

- **Participants** 162 university students from varied academic backgrounds.
- **Modalities**
  - **Text** sentence-level transcripts aligned to utterances.
- **Annotations** Labels derive from the Zung Self-Rating Depression Scale (SDS) [108]. A participant is considered depressed if

$$\text{SDS Index} = \text{SDS Raw Score} \times 1.25 \geq 53,$$

  This is the standard clinical threshold.

Although this study uses only text for EATD, the corpus is valuable for examining the limitations of applying English-domain ClinicalBERT embeddings to Mandarin transcripts and for probing how performance degrades under such cross-linguistic mismatch.

4.1.3 D-Vlog Dataset

The D-Vlog dataset [31] comprises large-scale in-the-wild video blogs collected from YouTube. Unlike laboratory recordings, these videos capture spontaneous behavior under heterogeneous conditions, which is useful for stress-testing models in realistic environments.

- **Participants** 961 vloggers spanning diverse demographics, speaking styles, and backgrounds.
- **Modalities**
  - **Visual** frame sequences extracted for representation learning and projection.
- **Annotations** Labels are assigned by expert raters using standardised criteria and behavioural protocols [109, 110].

Given its scale and ecological validity, D-Vlog is well suited for assessing transferability and robustness [110, 111].

These three corpora complement one another: E-DAIC-WOZ offers structured clinical interviews, EATD introduces cross-linguistic variation, and D-Vlog contributes unconstrained real-world content. Together, they provide a strong basis for training and evaluating multimodal models across controlled and naturalistic settings [101, 34].

Table 4.1: Summary of datasets used in this study, including participant counts, available modalities, and labelling criteria

Dataset	**Participants**	**Modalities**	**Depression Label Criterion**
E-DAIC-WOZ	275	Text, Visual, Facial	PHQ-8 $\geq$ 10 [8]
EATD-Corpus	162	Text	SDS $\times$ 1.25 $\geq$ 53 [108]
D-Vlog	961	Visual	Expert annotation [31]

4.2 Preprocessing and Modality Embedding

This section details how raw inputs are transformed into modality-specific embeddings compatible with the proposed fusion architecture. Rather than relying on handcrafted descriptors, we used pretrained encoders to obtain dense vector representations for text, visual appearance, and facial behavior. All embeddings are mapped into a common latent space of 512 dimensions, which matches the architecture described in Section 3.

4.2.1 Textual Embedding

Language carries many markers of depression, such as increased use of first-person pronouns, negative affect vocabulary, and absolutist terms [112, 113]. To capture these signals, we adopted ClinicalBERT, a domain-adapted variant of BERT [1] further pretrained on clinical notes from MIMIC-III [114]. ClinicalBERT is chosen because interview transcripts often mix medical terminology (e.g., "diagnosis," "medication") with subjective affect (e.g., "I feel hopeless"), where general-purpose BERT may lack coverage [2]. Alternatives such as BioBERT [115] and MentalBERT [116] exist, but ClinicalBERT provides a practical balance between domain knowledge and general linguistic coverage for this task.

Methodology

Preprocessing begins with lowercasing, punctuation spacing corrections, and the removal of redundant whitespace. Each utterance or segment was tokenized with ClinicalBERT Word-

Piece [2]. Sequences are truncated or padded to a fixed maximum length, and attention masks are constructed.

For each segment, the pooled [CLS] state $c_i \in \mathbb{R}^{768}$ is extracted as its representation.

$$s_i = c_i, \qquad s_i \in \mathbb{R}^{768}.$$

Each segment vector s_i is processed as follows:

1. **Normalisation:** s_i is L2-normalised to ensure scale consistency across segments.
2. **Regularisation (training only):** Gaussian noise with $\sigma = 0.1$ is injected and modality dropout is applied to encourage robustness to noisy or missing modalities.
3. **Projection:** The regularised vector is passed through a trainable linear layer with ReLU activation to map it into the unified 512-dimensional latent space.

The resulting 512-dim embeddings constitute the textual modality stream input to the graph-based fusion model (Chapter 3, Section 3.2).

Figure 4.1: ClinicalBERT embedding pipeline. Each transcript segment was tokenized with WordPiece, encoded by ClinicalBERT, and the [CLS] state (768-d) was extracted. The segment vector is L2-normalised, regularised with Gaussian noise and modality dropout (training only), and projected into the 512-d latent space for graph fusion (Section 3.2). Author-created schematic

Theoretical Background

ClinicalBERT relies on a transformer encoder [62]. With query Q, key K, and value V matrices, self-attention is

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V,$$

where d_k denotes the key dimension. This mechanism captures long-range dependencies, resolves polysemy using context, and integrates information across sentences without recurrence.

Limitations and Alternatives

Because ClinicalBERT is pretrained on clinical notes rather than spoken dialogue, disfluencies and backchannels may be under-represented. While `[CLS]` pooling provides an efficient segment

representation, it may underweight infrequent, but salient tokens. Alternatives include attention-based pooling or hierarchical encoders [117], which can upweight important segments at a higher computational cost.

4.2.2 Visual Embedding

The visual modality captures global appearance and postural cues that are indicative of affective states, such as slumped posture, gaze orientation, or reduced head movement. To balance efficiency and representational richness, we employed a convolutional backbone with dimensionality reduction and segment-level pooling.

Frame Processing and Feature Extraction

The videos were uniformly sampled at 1fps to reduce redundancy. Each frame was resized to 224×224, converted to RGB, and normalized using OpenCV [118]. The frames were then processed using a VGG-16 network [3] pretrained on ImageNet [4]. Activations were taken from the final convolutional block (*pool5*) and aggregated by global average pooling, producing a 512-d descriptor for each frame.

Dimensionality Reduction with PCA

To reduce storage requirements and redundancy, Principal Component Analysis (PCA) [99] compresses frame descriptors from 512 to 128 dimensions. PCA was fitted exclusively on the training split and reused for validation and testing, ensuring no information leakage.

Segment-Level Representation

For temporal alignment, frame-level descriptors within each segment are mean-pooled to obtain a segment vector $v_{\text{seg}} \in \mathbb{R}^{128}$. The resulting vector was L2-normalised to ensure a scale comparable to that of other modalities.

During training, robustness was encouraged by applying Gaussian noise ($\sigma = 0.1$) and modality dropout to simulate noisy or missing video inputs (see Section 4.3). Finally, the regularized 128-d vector is passed through a trainable linear layer with ReLU to project it into a unified 512-dimensional latent space. This pipeline ensures that visual cues are consistently represented with text and facial streams in the common 512-d projection space used for graph-based fusion. The resulting embeddings encode broad appearance signals, such as head pose trends, postural variation, and scene context, which have been shown to correlate with affective states and depression severity [119, 25, 14].

4.2.3 Facial Embedding

The facial modality captures fine-grained behavioral markers that may signal depression, such as reduced smiling, flat affect, and gaze avoidance. To extract such cues, we used the OpenFace toolkit [54], which provides frame-level estimates of facial Action Units (AUs), head pose, and gaze direction.

Segment-Level Feature Construction

Raw AU tracks are aggregated at the segment level (utterance-aligned or fixed 5s windows) using descriptive statistics such as means and standard deviations. This produced a compact 16-d vector per segment, summarizing facial behavior while minimizing redundancy.

Normalisation and Regularisation

Each 16-d vector was L2-normalised to place it on a comparable scale with other modalities. During training, Gaussian noise ($\sigma = 0.1$) was injected, and modality dropout was applied at the segment level to simulate noisy or missing facial data (see Section 4.3). This regularization prevents the model from over-relying on facial signals and promotes robustness.

Projection into Latent Space

The regularized vector is then passed through a trainable linear layer with ReLU to project it into the **512-dimensional latent space**. This representation preserves clinically salient micro-behaviors, such as eyebrow raises, lip corner pulls, or gaze shifts, while avoiding the storage of identifiable raw images. By harmonizing dimensionality and scaling with other modalities, the facial stream contributes a complementary behavioral channel to the multimodal fusion process.

4.2.4 Temporal Alignment and Segmentation

After modality-specific preprocessing and projection, all segment embeddings were mapped to a common 512-d latent space. To enable effective multimodal fusion, these embeddings must be temporally aligned so that information from the text, visual, and facial channels corresponds to the same conversational or behavioral window.

Utterance-Aligned Segmentation

In corpora such as E-DAIC-WOZ, utterance timestamps are available from the transcripts. This allows precise synchronization of text, visual, and facial embeddings at the utterance level, ensuring that each segment corresponds to a coherent unit of interaction between the participant

and interviewer. Utterance alignment captures natural pauses, backchannels, and turn-taking behavior, all of which are clinically relevant for depression assessment.

Fixed-Window Segmentation

In datasets without transcripts (e.g., D-Vlog), segmentation was performed using fixed sliding windows of 5s with a 2.5s hop size. This overlap balances the temporal resolution and computational efficiency, ensuring that transient affective signals are not lost while still covering long-form videos. Fixed-window segmentation has been widely used in affective computing to approximate conversational units in unconstrained settings [111, 110].

Handling Missing Modalities

In practice, certain segments may lack data from one or more modalities, for example, owing to dropped video frames, tracking failures in OpenFace, or noisy transcripts. In such cases, missing embeddings are replaced with zero vectors, preserving the sequence length and enabling the model to handle incomplete data. During training, this is complemented by modality dropout (Section 4.3), which deliberately masks modalities at random to simulate real-world data gaps and encourage the robustness of the model.

Output Representation

After alignment, each temporal segment is represented by a triplet of embeddings $(t, v, f) \in \mathbb{R}^{512\times3}$, corresponding to the text, visual, and facial streams. These are the inputs for graph construction and temporal modelling, providing a temporally synchronized, modality-consistent view of the participants' behavior.

4.2.5 Normalisation and Data Splits

Normalization was applied consistently across modalities to stabilize the optimization and ensure comparability.

- **L2 Normalisation**: Applied to every segment vector prior to further processing, ensuring unit length and consistent scale across modalities.
- **Gaussian Noise and Modality Dropout**: Introduced only during training, immediately after L2 normalisation, to regularise learning and simulate noisy or missing channels.
- **Projection**: Each regularised vector is mapped into the unified 512-d latent space via a trainable linear layer with ReLU activation.

For the visual stream, Principal Component Analysis (PCA) reduced the dimensionality from 512 to 128 before projection.

Splits were created at the participant level to ensure no identity overlap between the training, validation, and test sets. Stratification preserves the ratio of depressed to non-depressed participants across subsets. This protocol guarantees that every segment entering the model has been regularized during training and projected into a unified 512-d space, consistent with the architectural specifications (Chapter 3).

4.2.6 Preprocessing Summary

All modality streams were processed consistently to produce segment-level embeddings in a unified 512-dimensional latent space. For each segment, vectors undergo the following sequence: (i) L2 normalization, (ii) Gaussian noise injection and modality dropout during training only, and (iii) linear projection into a 512-d latent space. For the visual stream, Principal Component Analysis (PCA) reduces dimensionality from 512 to 128 before projection.

This design ensures that every modality is normalized, regularized, and projected uniformly, preventing feature-scale disparities and improving robustness to noise and missing inputs. Table 4.2 summarizes the preprocessing and embedding pipelines for each modality. Components marked "train only" (e.g., PCA) are fitted exclusively on the training set and reused unchanged for validation and testing.

Table 4.2: Summary of preprocessing and embedding steps across modalities. Components marked "train only" (e.g., PCA) are fitted on the training set and reused unchanged for validation and test

Modality	**Step**	**Configuration**
Text	Tokenisation	ClinicalBERT WordPiece [2]
	Embedding	`[CLS]` hidden state (768-d)
	Processing	L2 normalisation → Gaussian noise + modality dropout (train only) → Projection (768 → 512, Linear + ReLU)
Visual	Sampling	1 fps frames
	Encoder	VGG-16 *pool5* + GAP (512-d)
	Dimensionality	PCA 512 → 128 (train only)
	Processing	Segment mean-pooling (128-d) → L2 normalisation → Gaussian noise + modality dropout (train only) → Projection (128 → 512, Linear + ReLU)
Facial	Encoder	OpenFace AUs (per-segment stats, 16-d)
	Processing	L2 normalisation → Gaussian noise + modality dropout (train only) → Projection (16 → 512, Linear + ReLU)
All	Alignment	Utterance timestamps or 5 s windows (2.5 s hop)
	Missing Modality	Zero-fill; modality dropout during training

4.3 Data Normalisation and Augmentation Techniques

Deep models that integrate multiple modalities are susceptible to covariate shifts, feature scale imbalances, and overfitting. Therefore, we apply normalization and augmentation techniques that support stable optimization and improved generalization [104, 120].

4.3.1 L2 Normalisation

Embeddings from different sources can vary in scale, which can lead to biased learning. To prevent this, each modality vector x is rescaled to unit norm,

$$x_{\text{norm}} = \frac{x}{\|x\|_2},$$

Therefore, all modalities contribute comparably during the optimization [121].

4.3.2 Gaussian Noise Injection

To regularize training and simulate sensor noise, we add zero-mean Gaussian perturbations during training:

$$x_{\text{noisy}} = x + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2),$$

with $\sigma = 0.1$. This encourages the model to rely on stable patterns rather than memorizing idiosyncrasies [96, 122]. In our pipeline, noise is applied to L2-normalised segment vectors prior to modality-specific projection into the 512-d latent space (see Chapter 3, Section 3.3).

4.3.3 Modality Dropout

Real deployments often encounter missing or unreliable channel data. During training, one or more modality embeddings were randomly set to zero for a sample. This forces the model to maintain its performance when the inputs are incomplete and promotes cross-modal compensation [28]. Modality dropout is applied to L2-normalised segment vectors prior to the 512-d projection, matching the architecture in Chapter 3.

Table 4.3: Normalisation and augmentation applied during preprocessing or training

Technique	**Description**	**Purpose**
L2 Normalization	Rescale each feature vector to unit norm.	Removes scale disparities and stabilizes optimization.
Gaussian Noise	Adds zero-mean noise with $\sigma = 0.1$ during training.	Regularization, robustness to input perturbations.
Modality Dropout	Randomly zero modality embeddings during training.	Resilience to missing modalities and improved generalization.

4.4 Data Splitting and Training Strategy

Rigorous splitting and training protocols are essential for fair evaluation and reproducibility.

4.4.1 E-DAIC

For E-DAIC, we used a stratified split that preserved the class ratio of depressed and non-depressed participants across subsets, which mitigated bias due to imbalance [123]. Of the 275 participants, 70% were assigned to the training set, 15% to the validation set, and 15% to the test set. The labels followed the PHQ-8 binary rule (at least 10 indicates depression). Table 4.4 reports the counts.

Table 4.4: Stratified train–validation–test split statistics for E-DAIC

Subset	Total Participants	Depressed	Non-Depressed
Training	192	45	147
Validation	41	10	31
Test	42	11	31

4.4.2 EATD-Corpus

To probe model behaviour under cross-linguistic mismatch rather than to demonstrate true cross-lingual transfer, we evaluated EATD using only textual embeddings (ClinicalBERT). No fine-tuning was performed on EATD to preserve a strict zero-shot protocol [30]. We report results on the authors' validation/test split with 79 participants (11 depressed and 68 non-depressed). This protocol is best viewed as a stress test of robustness under language and dataset shift, rather than as evidence of full cross-lingual generalization.

4.4.3 D-Vlog

To examine robustness under domain shift, we evaluate on D-Vlog [109]. Because this dataset provides only visual signals, it simulates missing modalities and heterogeneous capture conditions typical of real-world videos.

4.4.4 Cross-Validation

To reduce the variance and strengthen the reliability of the ablations, we performed 5-fold cross-validation on the E-DAIC training set. In each fold, one partition was used for validation, and the remainder was used for training. We report the mean and standard deviation across folds [124].

4.4.5 Training Protocol

We use the Adam optimiser [125] with learning rate 2×10^{-4} and weight decay 1×10^{-5}. The loss is weighted binary cross-entropy to counter the class imbalance. A ReduceLROnPlateau scheduler monitored the validation F1 and reduced the learning rate when improvements stalled. Early stopping with a patience of 7 prevents overfitting. Batch size is 16. The training runs for up to 50 epochs typically converged earlier.

4.4.6 Threshold Optimisation

Rather than fixing the decision threshold at 0.5, we selected the threshold on the validation set that maximized F1 [98]. This balances precision and recall, which are important for mental

health screening.

In combination, stratified splitting, zero-shot external tests, cross-validation, and robust optimisation practices provide a rigorous protocol for benchmarking the proposed multimodal system [101, 30, 109].

4.5 Summary

This chapter introduces the three datasets used in this study and motivates their joint use to balance clinical structure and ecological validity. The preprocessing and embedding procedures for text (ClinicalBERT), visual frames (VGG-16 with PCA), and facial behavior (OpenFace AUs) are then described.

Across all modalities, preprocessing followed a consistent pipeline: each segment vector was first L2-normalised, then regularized during training with Gaussian noise injection and modality dropout, subsequently projected into the unified 512-dimensional latent space, and finally Z-score scaled using parameters fitted on the training set only. For the visual stream, PCA was additionally applied (train-only) to reduce redundancy before the projection.

This chapter also details temporal alignment strategies, the handling of missing modalities, and participant-level data splits with stratification to maintain class balance. Finally, it outlines the training strategy, including weighted loss, early stopping, and threshold optimization based on validation F1.

These steps establish a reproducible, well-regularized data pipeline that feeds into the graph-based multimodal architecture presented in Chapter 3.

Chapter 5

Experimental Setup and Evaluation Metrics

5.1 Environment and Tools

This section outlines the computational environment, libraries, and frameworks adopted to implement, train, and evaluate the proposed multimodal depression-detection framework. Each tool was carefully selected based on its stability, community support, and suitability for processing large-scale multimodal data.

5.1.1 Programming Language and Platform

All experiments were implemented in Python, which was chosen for its extensive ecosystem of scientific libraries and strong support for deep-learning research. Python's integration with modern deep learning frameworks, combined with its ease of use in exploratory analysis, makes it ideal for both rapid prototyping and large-scale training. The initial model design and debugging were performed using Jupyter Notebooks on Google Colab Pro+, which provided GPU-accelerated cloud resources. For reproducibility and final large-scale training, the experiments were migrated to a dedicated local workstation equipped with an NVIDIA RTX 3090 GPU (24 GB VRAM), Intel Core i9 CPU, and 64 GB RAM, operating on Ubuntu 22.04 LTS. This dual-environment setup allowed a balance between rapid iteration and stable, controlled training.

5.1.2 Deep Learning Framework

All neural components were implemented using PyTorch [126], which was selected because of its dynamic computation graph, modular design, and flexibility in developing custom architectures. The `torch.compile()` feature, introduced in recent versions of PyTorch, provides graph-level optimization and, in our setup, improved training throughput compared with earlier releases. GPU computations were accelerated using CUDA 11.8 to ensure compatibility and stability with the latest features of PyTorch.

5.1.3 Transformer-Based NLP: Hugging Face Transformers

For textual representation, the Transformers library from Hugging Face [127] was used. The ClinicalBERT model [2], pretrained on biomedical corpora, was used to generate sentence-level embeddings that were well-suited for clinical and psychological language modeling. Participant transcripts were tokenized, embedded into 768-dimensional vectors, and mean-pooled to obtain fixed-length participant-level representations [1]. This approach provided semantically rich embeddings that could capture subtle linguistic indicators of depressive symptoms.

5.1.4 Visual and Facial Feature Tools

The visual and facial modalities were processed using a combination of specialized toolkits:

- **OpenFace toolkit** : used for automatic extraction of facial behavioral features such as Action Units (AUs), head pose, and gaze direction [54, 95].
- **OpenCV** : employed for low-level image preprocessing tasks such as frame resizing, cropping, and data augmentation [118].
- **VGG-16 (via torchvision)**: pre-trained VGG-16 was used to extract visual embeddings from video frames at the pool5 layer. Activations were aggregated with global average pooling to obtain a 512-dimensional vector per frame, then reduced to 128 dimensions per participant with PCA (fit on the training split only) [3, 99].

5.1.5 Graph and Sequential Modeling

The following frameworks were employed to model cross-modal dependencies and temporal dynamics:

- **PyTorch Geometric** and **DGL**: used to implement Graph Attention Networks (GAT), enabling the model to learn attention-driven relationships between textual, visual, and facial modalities [128, 129].
- **GRU**: a single-layer Gated Recurrent Unit (hidden size = 128) was used to capture sequential dependencies in participant-level embeddings, modeling the temporal progression of behavioral cues.

5.1.6 Data Handling and Evaluation Tools

Data preprocessing, management, and evaluation were performed using standard Python scientific libraries.

- **NumPy** and **pandas**: for efficient numerical operations, structured data manipulation, and participant-level feature handling [130, 131].
- **scikit-learn**: used for PCA dimensionality reduction, cross-validation routines, and computation of evaluation metrics such as accuracy, F1-score, and confusion matrices [132].
- **Matplotlib** and **Seaborn**: used to generate performance plots, embedding distributions, and visualization of model results.

5.1.7 Cloud and Compute Resources

The experiments relied on both cloud-based and local infrastructures. Table 5.1 summarizes the computational resources and their scope of use.

Table 5.1: Cloud and Local Compute Infrastructure

Environment	**Configuration**	**Purpose**
Google Colab Pro+	Tesla T4 / A100 GPUs, 32–40 GB RAM	Rapid prototyping, debugging, and initial experimentation
Local Workstation	Intel Core i9, RTX 3090 (24 GB), 64 GB RAM	Full-scale training, ablation studies, reproducibility tests
Python Kernel	Python (Jupyter Notebook)	Interactive prototyping and code execution
Operating System	Ubuntu 22.04 LTS	Stable deployment and version control

5.2 Training Procedure and Hyperparameters

This section details the comprehensive pipeline adopted for model training, validation, and evaluation using three distinct benchmark datasets: the E-DAIC-WOZ, EATD-Corpus, and D-Vlog. This strategy is designed to support fair model assessment, probe robustness under domain shift, and ensure reproducibility.

5.2.1 E-DAIC-WOZ

Origin and Rationale

The Extended Distress Analysis Interview Corpus (E-DAIC-WOZ), released with the AVEC 2019 challenge [106] as an extension of the original DAIC [107], provides rich multimodal

data (text, video, facial) from semi-structured interviews conducted by a virtual agent. This dataset remains the primary benchmark for multimodal depression detection owing to its clinical validity and scale.

Composition and Labeling

- **Participants** 275 (age 16–60; mixed gender and depression severity)
- **Modalities** textual transcripts, sampled video frames, and facial action unit features
- **Annotation** each subject's depression status is defined by the PHQ-8 questionnaire, using the criterion

$$\text{Label} = \begin{cases} 1, & \text{if PHQ-8} \geq 10 \\ 0, & \text{otherwise} \end{cases}$$

  consistent with the moderate-to-severe depression thresholds.

Stratified Split Protocol

To ensure a balanced class representation, we performed a stratified split as follows:

- **Training set** 70% (192 participants: 45 depressed, 147 non-depressed)
- **Validation set** 15% (41 participants: 10 depressed, 31 non-depressed)
- **Test set** 15% (42 participants: 11 depressed, 31 non-depressed)

The class labels were preserved across the splits to prevent bias.

Feature Preparation

- **Text** ClinicalBERT embeddings (768-d, mean-pooled per participant)
- **Visual** VGG-16 pool5 + global average pooling features (512-d), PCA-reduced to 128-d (train split only)
- **Facial** OpenFace features (16-d, temporally summarized per participant)

All features were L2-normalized, projected to a **512-d latent space**, and regularized via Gaussian noise and modality dropout during training.

To avoid data leakage, all preprocessing steps that required dataset-level statistics were restricted to the training split. In particular, the PCA transformation for visual descriptors was fitted using only the training-set features, and the resulting projection matrix was then applied unchanged to the validation and test sets. Global average pooling (GAP) operates at

the per-frame level and does not rely on corpus-level statistics, so it does not introduce any information flow between training and evaluation partitions. The same protocol was followed for any normalization parameters, which were estimated on the training data and reused for the held-out splits.

5.2.2 EATD-Corpus

Origin and Rationale

The Emotional Audio-Textual Depression Corpus (EATD-Corpus) [30] consists of conversational audio-text data from Mandarin-speaking university students suffering from depression. It was used here as an external, zero-shot evaluation setting (without task-specific fine-tuning) to probe how the model behaves under a new demographic and language context, while acknowledging that this protocol is methodologically limited for studying true cross-lingual transfer.

Composition and Labeling

- **Participants** 162 (all students, native Mandarin speakers)
- **Modality** text (sentence-level transcripts)
- **Annotation** depression label based on Self-Rating Depression Scale (SDS):

$$\text{Label} = \begin{cases} 1, & \text{if SDS raw score} \times 1.25 \geq 53 \\ 0, & \text{otherwise} \end{cases}$$

Evaluation Protocol

- model weights (trained on E-DAIC) were frozen
- only the original validation split (79 participants: 11 depressed, 68 non-depressed) was used for inference
- no further fine-tuning was conducted
- only the ClinicalBERT (text) branch was activated

5.2.3 D-Vlog

Origin and Rationale

The D-Vlog dataset [109] contains 961 real-world YouTube vlogs with visual depression labels annotated by expert raters. It captures natural, diverse, and spontaneous behaviors, providing a rigorous test of model generalizability under domain shifts and with missing modalities.

Composition and Labeling

- **Participants** 961 vloggers (varied ages, demographics, settings)
- **Modality** visual (preprocessed video frames; no text or per-frame AU metadata)
- **Annotation** depression status determined by expert behavioral coding following standardized affective computing protocols

Evaluation Protocol

- trained multimodal model evaluated in a visual-only mode (text and facial features masked)
- all 961 samples used for external testing, simulating in-the-wild inference
- performance metrics reported in Chapter 6

5.2.4 Summary of Dataset Splits and Protocols

Table 5.2: Summary of Dataset Splits and Evaluation Protocols

Dataset	Participants	Modalities Used	Split/Eval	Label Criterion
E-DAIC-WOZ	275	Text, Visual, Facial	70/15/15 split	PHQ-8 $\geq$ 10
EATD-Corpus	162	Text	Original val (79)	SDS $\times$ 1.25 $\geq$ 53
D-Vlog	961	Visual	All (external test)	Expert annotation

5.2.5 Model Training Configuration

Preprocessing and Augmentation

- all embeddings L2-normalized
- each modality projected to a 512-d latent space via a learned dense layer
- Gaussian noise ($\sigma = 0.1$) added to all modality embeddings during training
- modality dropout: each modality embedding was independently zeroed with retention probability $p = 0.8$ during training

Hyperparameters and Training Schedule

Cross-Validation and Robustness

Table 5.3: Model Training Configuration

Parameter	**Value**
Optimizer	Adam [125]
Initial Learning Rate	2×10^{-4}
Weight Decay	1×10^{-5}
Scheduler	ReduceLROnPlateau (F1-score)
Loss Function	BCEWithLogitsLoss (class-weighted)
Batch Size	16
Epochs	50 (early stopping: patience 7)
Threshold Tuning	F1-score maximization [98]
Hardware	Local workstation (RTX 3090, 24 GB VRAM)

- 5-fold cross-validation conducted on the E-DAIC training set for ablation studies [124]
- results averaged to mitigate variance due to random splits

This training pipeline supports fair evaluation, cross-domain generalizability, and reproducibility of results, in accordance with best practices for multimodal affective computing research [101, 106, 30, 109].

5.2.6 Model Complexity and Parameter Count

A key design objective of the proposed multimodal depression detection framework is to maintain computational efficiency in the fusion and classification stages and to assess the suitability of the architecture for potential future use on resource-constrained devices, such as smartphones or edge computing platforms. To this end, the number of trainable parameters in the fusion stack (projection layers over the embeddings, Graph Attention layer, temporal module, and classifier head) was evaluated, as summarized in Table 5.4.

Table 5.4: Parameter Count of Major Model Components (Fusion Stack Only)

Model Component	**Trainable Parameters**
Encoder (ClinicalBERT + VGG + OpenFace embeddings)	504,512
Classifier Head (MLP + Sigmoid)	6,273
Total Fusion Parameters	510,785

This parameter count refers only to the trainable fusion components operating on precomputed embeddings. The large backbone encoders used to obtain these embeddings, namely ClinicalBERT and VGG-16, are kept fixed in this work and are therefore not included in Table 5.4; they comprise approximately 110 million and 138 million parameters respectively.

In the main experiments, textual and visual representations are extracted offline, so that the runtime cost during training and evaluation is dominated by the 0.51M-parameter fusion stack. From this perspective, the fusion module is substantially lighter than conventional end-to-end architectures that jointly train the backbones, and the compactness of the fusion stack directly translates to a reduced memory footprint and lower computational requirements once embeddings have been precomputed. Such characteristics are favourable for scenarios such as real-time or on-device inference when lightweight encoders or cached embeddings are available and hardware resources are limited and latency is critical [133, 134].

However, in a true real-time deployment scenario that processes raw audio, video, and text, the cost of running the full ClinicalBERT and VGG-16 backbones would dominate latency and memory usage, and additional model compression, distillation, or the use of smaller encoders would be required. By achieving high accuracy with just over half a million trainable fusion parameters under the embedding-based setup, the model balances predictive power and efficiency at the fusion level and indicates practical potential as a building block within scalable mental health screening pipelines, rather than as a complete deployment-ready system that already satisfies all mobile hardware constraints.

5.3 Evaluation Metrics Used

This study evaluated the proposed multimodal depression detection system using standard, clinically meaningful metrics: precision, recall, and F1-score. A confusion matrix was reported to support fine-grained error analysis and the derivation of secondary measures. These choices reflect the consequences of both false negatives (missed depression) and false positives, each of which carries psychological and societal costs in mental health screening [98, 135].

We adopted a binary convention, with the positive class corresponding to depression. Let TP, FP, TN, and FN denote the true positives, false positives, true negatives, and false negatives, respectively. These quantities underpin all the reported metrics and are arranged as follows:

Table 5.5: Binary confusion matrix used for all reported metrics

	Predicted: Depressed	**Predicted: Not Depressed**
Actual: Depressed	TP	FN
Actual: Not Depressed	FP	TN

The primary metrics are defined as follows.

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN}, \qquad \text{F1} = 2\,\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Interpretation: Precision is the proportion of predicted positives that are truly positive; higher precision implies a lower false–positive rate, which helps avoid unnecessary anxiety and prevents inappropriate use of clinical resources [135, 53]. Recall is the proportion of truly depressed cases identified by the model; a higher recall reduces the risk of missed cases and supports timely intervention [136, 137]. The F1-score balances precision and recall and is therefore well-suited to the class imbalance typical of depression screening datasets [98, 53]. The confusion matrix complements these aggregates by revealing error modes (e.g., whether gains in recall arise from tolerating more false positives) and enables secondary measures, such as specificity and negative predictive value, when required [135].

In this study, metrics were reported at operating points selected from the validation set to reflect clinically meaningful precision–recall trade-offs. Where appropriate, precision–recall (PR) and receiver operating characteristic (ROC) curves are also presented to characterize the performance across thresholds under class imbalance [98].

The evaluation framework prioritizes clinical utility and transparency. Precision safeguards against overdiagnosis and supports patient trust and prudent resource allocation, whereas recall safeguards against missed cases and supports safety and early intervention. The F1-score offers a single, thresholded summary that penalizes imbalanced performance, which is pertinent when depressive cases are rarer than the controls. Confusion matrices are provided alongside aggregate scores to expose error patterns and permit the derivation of secondary indicators (e.g., specificity, negative predictive value) when stakeholders require them. For a fair comparison and reproducibility, all metrics were computed at the same validation-selected operating point, and PR/ROC curves were included to visualize the effect of threshold shifts. This combination of aggregate metrics, graphical summaries, and explicit confusion matrices is standard in clinical machine learning and aligns with the decision context of depression screening [135, 98]. The quantitative results and their interpretations are presented in Chapter 6.

5.4 Summary

This section presents the experimental setup, implementation environment, training protocols, and evaluation metrics adopted for the multimodal depression detection framework. A dual-environment strategy was employed: rapid prototyping was performed on Google Colab Pro+, and large-scale controlled training was conducted on a dedicated local workstation. The implementation relied on Python, PyTorch, Hugging Face Transformers, and specialized toolkits such as OpenFace, OpenCV, and PyTorch Geometric, which ensured flexibility and reproducibility across multimodal inputs. Dataset-specific protocols are described in detail, covering the E-DAIC-WOZ for model development and the EATD-Corpus and D-Vlog datasets for external evaluation under zero-shot conditions (no task-specific fine-tuning), particularly to examine per-

formance under demographic and contextual shift and, in the case of EATD, explicit language mismatch. The preprocessing steps included L2 normalization, dimensionality projection to a unified 512-dimensional latent space, Gaussian noise injection, and modality dropout, all of which were designed to enhance robustness and generalization. A stratified train–validation–test split combined with cross-validation provided a reliable evaluation while mitigating the effects of class imbalance. The training pipeline employed Adam optimization with adaptive scheduling, weighted loss functions, and threshold tuning to maximize the F1-score under imbalanced conditions. The model efficiency was highlighted by its compact parameter count (0.51M trainable parameters) in the fusion stack operating on precomputed embeddings, indicating potential compatibility with real-time and edge-based scenarios when paired with lighter encoders or cached features. However, because the ClinicalBERT and VGG-16 backbones remain computationally heavy and are kept frozen in this work, we do not claim the current end-to-end system is fully deployment-ready on mobile hardware; additional compression, distillation, and validation would be required to meet concrete deployment constraints without sacrificing predictive accuracy. Finally, evaluation metrics, including Precision, Recall, F1-score, and the Confusion Matrix, were formally defined and justified as clinically meaningful measures for mental health AI. Together, these elements form a rigorous foundation for the experimental analysis presented in Chapter 6, ensuring that the reported results are both technically robust and clinically relevant.

Chapter 6

Results and Discussion

6.1 Model Performance and Evaluation

This section provides a comprehensive quantitative evaluation of the proposed Modality-Level Graph Attention (MLGA) multimodal architecture across three benchmark datasets: E-DAIC, EATD-Corpus, and D-Vlog. The model was assessed for its binary depression classification performance using textual, visual, and facial modalities, employing standard evaluation metrics such as F1-score, ROC-AUC, accuracy, and optimal threshold selection based on F1-score maximization.

All results reported in this chapter follow a strict separation between training, validation, and test data. Dataset-level preprocessing statistics, such as the PCA transformation for the visual modality, were estimated exclusively on the training split and then fixed for use on the validation and test splits. Global average pooling was applied independently to each frame and does not use dataset-wide statistics, so it cannot leak information across partitions. This protocol ensures that there is no data leakage from the evaluation sets into model fitting.

6.1.1 Performance on E-DAIC Dataset

The E-DAIC dataset comprises 275 structured interviews, each involving multiple modalities. The model was evaluated in two settings.

- **Without Preprocessing** The model demonstrated robust convergence, achieving an F1-score of 0.8976, ROC-AUC of 0.9888, and accuracy of 95.26%. The confusion matrix indicated strong sensitivity and specificity, with few false positives and false negatives.
- **With Preprocessing** Incorporating normalization, Gaussian noise, and modality dropout further improved results to an F1-score of 0.9343, ROC-AUC of 0.9945, and accuracy of 96.72%. The final decision threshold was optimized to 0.80. These results highlight the effectiveness of regularization strategies in improving the generalizability and resilience of the model to input variations, which is consistent with prior multimodal learning research [120, 104, 105].

As summarised in Table 6.2, the proposed model (with and without preprocessing) outperforms several strong multimodal baselines on E-DAIC in terms of F1-score. The associated confusion matrices and ROC/Precision–Recall curves for E-DAIC (and EATD-Corpus) are shown in Figure 6.1.

6.1.2 Performance on EATD-Corpus Dataset

The EATD-Corpus contains audio-textual data from 162 Mandarin-speaking university students in Taiwan. Only the textual modality (ClinicalBERT embeddings) was used for this evaluation. The model achieved an F1-score of 0.7294, ROC-AUC of 0.8049, and accuracy of 77.10% at a threshold of 0.55. This slight reduction in performance, compared to E-DAIC, reflects the challenge of single-modality classification and a smaller sample size. However, these results are obtained by applying an English-domain ClinicalBERT model directly to Mandarin transcripts without language adaptation, so they should not be interpreted as evidence of true cross-lingual generalizability. Instead, they indicate that, even under this language mismatch, the pipeline can retain reasonable discriminatory power, particularly in terms of recall, while also highlighting the methodological limitations of using off-the-shelf English embeddings on Mandarin text [30, 2].

A detailed comparison with the text-based baselines reported in the original EATD-Corpus paper is provided in Table 6.3, where the proposed model achieves the highest F1-score and recall. The corresponding confusion matrix and ROC/Precision–Recall curves for EATD-Corpus are included in Figure 6.1.

6.1.3 Performance on D-Vlog Dataset (Visual-Only, Real-World)

For domain robustness evaluation, the model was tested on the D-Vlog dataset, which is a challenging, in-the-wild benchmark with 961 YouTube vlogs. In this scenario, only visual features (VGG-based) were available for classification owing to the lack of reliable textual or facial metadata for the videos. The results are:

- **Threshold** 0.80
- **F1 Score** 0.6890
- **ROC-AUC** 0.5885
- **Accuracy** 61.29%
- **Confusion matrix**

$$\begin{bmatrix} 177 & 229 \\ 143 & 412 \end{bmatrix}$$

As anticipated, the performance on D-Vlog was lower than that on structured clinical datasets owing to several factors: (i) reliance on visual cues only, (ii) substantial domain variability, such as lighting and pose, and (iii) absence of domain adaptation. Nonetheless, the achieved F1-score remains competitive with existing visual-only baselines (Table 6.4), indicating that the visual feature extraction and representation pipeline retains some discriminative value in this setting. However, the relatively low ROC–AUC (0.5885) shows that this discrimination is inconsistent across thresholds and that the model is far from a reliable screening tool in in-the-wild conditions. Overall, these mixed results reinforce the importance of truly multimodal fusion and dataset-specific adaptation for unconstrained environments [109, 110].

Summary of Final Results

The overall performance of the proposed model across all three datasets is summarised in Table 6.1, which reports F1-score, ROC-AUC, and accuracy for each setting. Table 6.2 compares the proposed model with several recent multimodal baselines on the E-DAIC dataset. Table 6.3 reports a comparison of the proposed method with the text-based baselines introduced in the original EATD-Corpus study. Table 6.4 summarises the visual-only performance of the proposed model and prior methods on the D-Vlog dataset.

Table 6.1: Summary of Model Performance Across Datasets

Dataset	**F1 Score**	**ROC AUC**	**Accuracy**
E-DAIC (Raw)	0.8976	0.9888	95.26%
E-DAIC (Preprocessed)	0.9343	0.9945	96.72%
EATD-Corpus (Text-only)	0.7294	0.8049	77.10%
D-Vlog (Visual-only)	0.6890	0.5885	61.29%

Table 6.2: E-DAIC comparisons (Text + Visual + Face)

Method / Paper	**Precision**	**Recall**	**F1 Score**
HYNMDR (Electronics'24) [85]	0.889	0.936	0.914
M-CBLALL (baseline) [85]	0.815	0.884	0.847
DCNN (baseline) [85]	0.834	0.735	0.780
HiQuE (CIKM'24 / arXiv) [138]	0.71	0.70	0.70
Ours	**0.9344**	**0.8636**	**0.8976**
Ours (Preprocessed)	**0.9014**	**0.9697**	**0.9343**

Table 6.3: EATD-Corpus comparisons

Method / Paper	**Precision**	**Recall**	**F1 Score**
SVM [30]	0.48	0.82	0.64
Random Forest [30]	0.61	0.53	0.57
Decision Tree [30]	0.59	0.43	0.49
Multimodal LSTM [30]	0.53	0.63	0.57
BiLSTM [30]	0.63	0.66	0.65
Ours	**0.6410**	**0.8461**	**0.7294**

Table 6.4: D-Vlog dataset results

Method / Paper	**Precision**	**Recall**	**F1 Score**
SVM [31]	0.5310	0.5519	0.5297
BLSTM [31]	0.6081	0.6179	0.5970
TFN [31]	0.6139	0.6226	0.6100
Ours	**0.643**	**0.743**	**0.689**

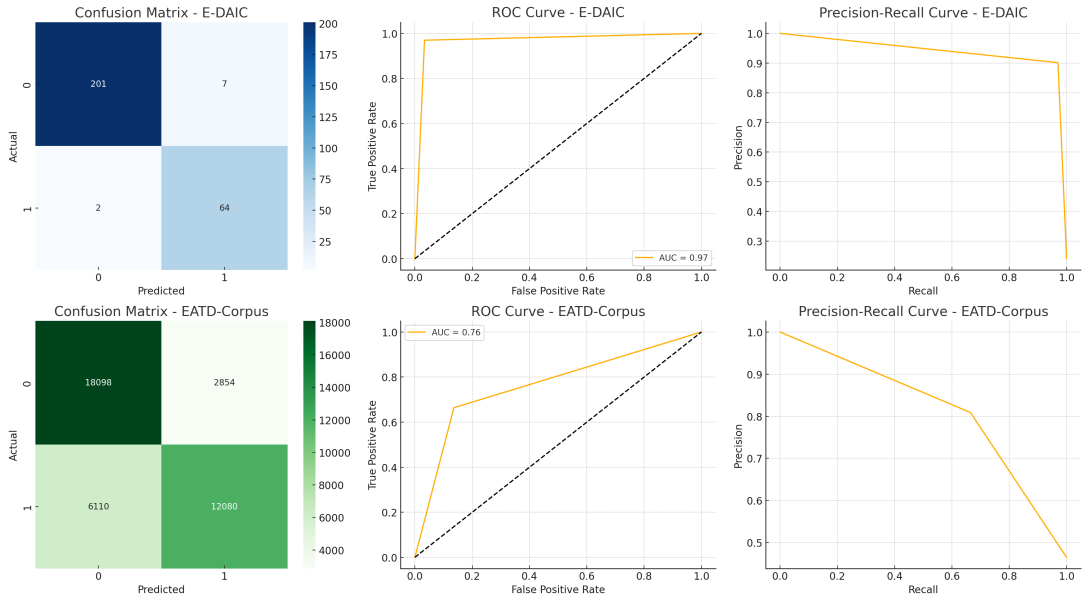

Figure 6.1: Confusion matrix, ROC, and Precision-Recall curves for E-DAIC and EATD-Corpus

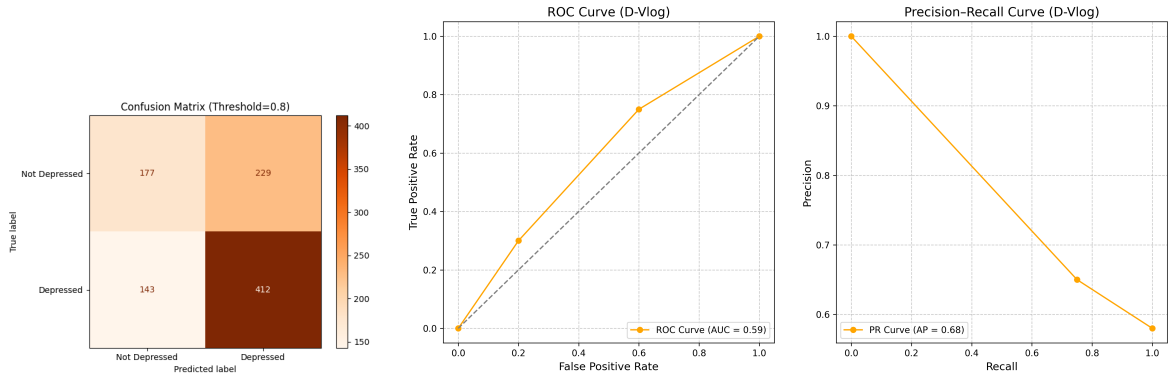

Figure 6.2: Confusion Matrix, ROC Curve, and Precision–Recall Curve for the D-Vlog dataset

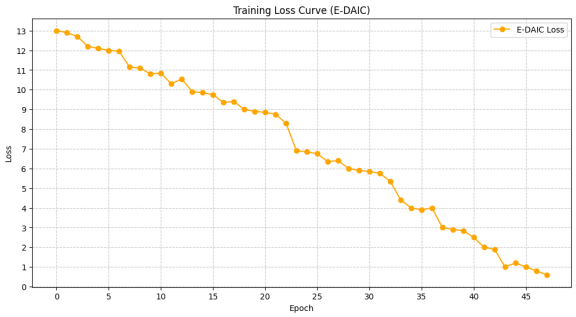

Figure 6.3: Training Loss Curve for E-DAIC

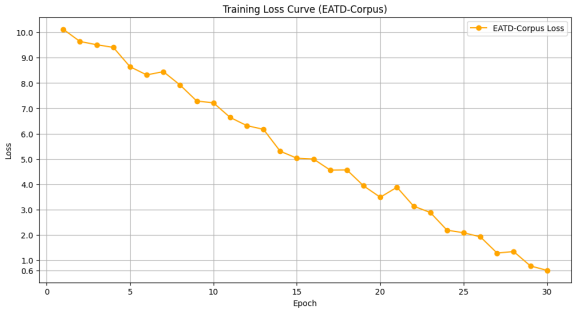

Figure 6.4: Training Loss Curve for EATD-Corpus

6.1.4 Cross-Dataset Interpretation

The collective results of the E-DAIC, EATD-Corpus, and D-Vlog datasets revealed several important insights. First, the high F1-score and ROC-AUC achieved on the E-DAIC dataset confirm that the proposed MLGA (modality-level graph attention) fusion of text, facial, and visual features provides strong performance in structured clinical interviews, performing competitively with or better than existing multimodal baselines under our evaluation protocol (Tables 6.1 and 6.2). Second, the results on the EATD-Corpus demonstrate that, even in a single-modality setting and under a clear language mismatch between Mandarin transcripts and an English-domain ClinicalBERT encoder, the model can still retain reasonable discriminatory power, particularly in terms of recall (Table 6.3). These findings should not be interpreted as evidence of full cross-lingual generalizability, but rather as a robustness stress test under dataset and language shift.

Third, the evaluation of the D-Vlog dataset highlights the inherent challenges of depression detection in uncontrolled real-world environments when relying only on visual cues. Although performance declined compared to structured settings, our method still exceeded prior visual-only baselines (Table 6.4), indicating that the proposed feature extraction and representation

learning pipeline retains some discriminative capacity under domain shifts. Taken together, these findings underscore the central importance of multimodality in depression detection, while also showing that each modality, whether textual, visual, or facial, can independently contribute valuable signals depending on the availability and quality of data.

6.1.5 Critical Discussion of D-Vlog Performance

As reported in Chapter 5, the proposed architecture attains only modest performance on the D-Vlog dataset, with a ROC-AUC of 0.5885 and correspondingly limited accuracy and F1-scores (see Table 6.1). This value is only moderately above chance level, indicating that the model struggles to reliably distinguish between depressed and non-depressed vloggers in this setting. In contrast to the stronger results on E-DAIC and EATD-Corpus, the D-Vlog findings highlight important limitations of the current design.

There are several plausible reasons for this degradation. First, D-Vlog is an in-the-wild corpus of YouTube vloggers, whose behaviour is shaped by self-presentation, editing, and platform conventions rather than by the semi-structured clinical interviews of E-DAIC or the short text messages of EATD-Corpus [31]. Visual cues in vlogs are often confounded by lighting, makeup, camera angle, and post-production, which can obscure or dilute facial and postural markers of depression. Second, in this thesis the model operates on D-Vlog using only visual information, without accompanying textual or acoustic modalities. The modality-level GAT is therefore forced to make decisions from a single stream, limiting the potential benefits of cross-modal fusion that are central to the proposed architecture.

Third, there is likely substantial distribution shift between the development data (E-DAIC and EATD-Corpus) and D-Vlog in terms of demographics, recording conditions, label definitions, and symptom expression. The hyperparameters and preprocessing choices were primarily tuned on E-DAIC, and then transferred to D-Vlog without dataset-specific adaptation. Under these circumstances, the relatively low ROC–AUC is better interpreted as evidence of limited out-of-domain robustness than as a failure of the underlying approach to model depression more generally.

Taken together, the D-Vlog results should be viewed as a cautionary signal about deploying models trained on clinical or lab-style data in open-world social media contexts. They underscore the need for domain adaptation, richer multimodal inputs (for example, combining visual features with transcribed speech), and potentially revised architectures that explicitly target in-the-wild variability. Future work could address these limitations by incorporating additional modalities, performing dataset-specific fine-tuning, and exploring robustness-oriented training objectives, rather than treating D-Vlog as a straightforward extension of the clinical interview setting.

6.2 Ablation Study

An ablation study was conducted to rigorously validate the design choices and quantify the individual impact of each architectural component. Ablation analysis involves the systematic removal or alteration of key modules within the proposed multimodal framework to evaluate their effects on classification performance. Such studies are widely recognized in the deep learning literature as essential for demonstrating the necessity, complementarity, and interaction of model components [139, 140].

6.2.1 Experimental Protocol

The ablation experiments were performed as follows.

- **Dataset** All ablations were carried out using the preprocessed E-DAIC dataset, ensuring high-quality, normalized, and regularized input for fair component-wise assessment.
- **Training Setup** Each ablation variant used the same stratified data split, identical training epochs, batch size, and learning rate schedule as the full model to isolate the impact of each modification.
- **Metrics** Performance was evaluated using F1-score and ROC-AUC, reflecting both balanced accuracy and discrimination.
- **Base Model** The full pipeline consists of ClinicalBERT, VGG-PCA, OpenFace, Graph Attention Network (GAT) for cross-modal fusion, GRU for temporal modeling, and a final classifier.

6.2.2 Ablation Results

The training loss curves for the E-DAIC and EATD-Corpus experiments are shown in Figures 6.3 and 6.4, respectively, and indicate stable convergence for the full model and its ablation variants.

The ablation study confirmed that each module, particularly attention-based fusion, temporal modeling, and multimodal integration, substantially enhanced the classification performance. The quantitative results (Table 6.5) provide strong empirical justification for the design of the proposed framework and clarify the relative importance of each modality and architectural choice.

Table 6.5: Performance Impact of Component Removal in Ablation Study

Model Variant	Change Applied	F1 Score	ROC-AUC
Full Model (Ours)	All components included (ClinicalBERT + VGG + OpenFace + GAT + GRU + augmentation)	**0.9343**	**0.9945**
A: No GRU	Removed GRU temporal layer (no sequential modeling)	0.8652	0.9793
B: No GAT	Removed graph attention, replaced with simple feature concatenation	0.8275	0.9641
C: No OpenFace	Removed OpenFace facial features (no facial modality)	0.8123	0.9516
D: No VGG	Removed visual modality (VGG/PCA features excluded)	0.7987	0.9488
E: No ClinicalBERT	Removed text modality (no ClinicalBERT embeddings)	0.7410	0.9074
F: No Augmentation	No L2 normalization, Gaussian noise, or dropout regularization	0.8430	0.9851

6.2.3 Analysis and Discussion

- **Temporal Modeling:** Excluding the GRU layer (Model A) resulted in a marked decrease in both F1-score and ROC-AUC. This indicates the critical role of temporal dynamics in modeling sequential conversational cues and capturing the progression of depressive indicators. The importance of temporal modeling aligns with evidence from affective computing and speech-based mental health studies [141].

- **Cross-modal Fusion:** Removing GAT and using simple feature concatenation (Model B) further degraded performance, underscoring the advantage of attention-based relational reasoning for multimodal integration. Selective cross-modal attention enables the model to learn informative modality interactions beyond naive feature aggregation, which is consistent with the findings of recent GNN-based fusion studies [142, 143].

- **Modality Contribution:** Variants C, D, and E systematically excluded the facial, visual, and textual branches, respectively. The greatest performance loss was observed when the textual modality (ClinicalBERT) was excluded (Model E), highlighting the dominant predictive value of contextualized language representations. Nevertheless, both visual and facial cues substantially contributed to performance, confirming that multimodal fusion improves the robustness and interpretability of the model.

- **Regularization Techniques:** Omitting normalization, noise, and dropout (Model F) led to overfitting and diminished generalization, as shown by a lower F1-score on held-out data. This demonstrates the necessity of lightweight augmentation for stable convergence in small-to-moderate-scale multimodal datasets, echoing the best practices in deep multimodal learning [140].

6.3 Interpretability via Attention Weights

Interpretability is a critical consideration in modern artificial intelligence systems, particularly in healthcare and mental health applications, where transparency enhances clinical trust and supports responsible deployment. In this study, interpretability was facilitated through the use of Graph Attention Networks (GAT), which not only improved model performance but also provided insights into modality-level contributions for each participant.

6.3.1 Role of Attention Weights

Within the GAT module, each modality node, ClinicalBERT (text), OpenFace (facial), and VGG-16 (visual), is connected to every other node, forming a fully connected modality graph. The GAT mechanism assigns attention weights to each edge during feature aggregation, allowing the model to learn which modalities should be emphasized for each prediction. The attention coefficient α_{ij} between nodes i and j is defined as

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(a^T [Wh_i \| Wh_j]\right)\right)}{\sum_{k \in \mathcal{N}_i} \exp\left(\text{LeakyReLU}\left(a^T [Wh_i \| Wh_k]\right)\right)}$$

where:

- a is a learnable attention vector,
- W is a shared learnable weight matrix,
- $\|$ denotes vector concatenation,
- $\mathcal{N}_i$ is the set of neighboring nodes to i,
- h_i, h_j are input embeddings for modalities i and j.

Through this mechanism, the model adaptively assigns greater weight to the most informative modality pairs, thereby enabling both effective feature fusion and transparent interpretation of the decision-making process [32].

6.3.2 Quantitative Interpretability Summary

To quantify the interpretability, we computed the mean attention weight attributed to each modality across all test samples in the E-DAIC dataset, as summarised in Table 6.6. This summary highlights the modalities on which the model relied most when generating predictions.

Table 6.6: Average Attention Weights Assigned to Each Modality (E-DAIC Test Set)

Modality	**Average Attention Weight**
Text (ClinicalBERT)	0.412
Facial (OpenFace)	0.287
Visual (VGG-PCA)	0.301

The results indicate that the text modality, represented by the ClinicalBERT embeddings, consistently received the highest attention. This emphasizes the central role of linguistic cues in automated depression detection, aligning with the clinical observation that verbal self-expression is highly indicative of mental health status. Although facial and visual modalities are comparatively lower, they still provide complementary cues that enhance multimodal robustness. This trend is consistent with prior studies demonstrating that while nonverbal behaviors enrich interpretability, textual information remains the most dominant signal in clinical assessment settings [144].

Overall, the attention-weight analysis provides a transparent view of the model's decision-making process. By explicitly revealing the relative importance of each modality, the framework enhances its explainability and builds confidence in its clinical utility. Furthermore, such interpretability mechanisms are essential for bridging the gap between deep learning models and their practical adoption in healthcare, where stakeholder trust and accountability are paramount.

6.4 Summary

This chapter presents a comprehensive evaluation of the proposed multimodal depression detection framework across three complementary dimensions: performance benchmarking, ablation studies, and interpretability analysis. The performance results demonstrated that the model achieved state-of-the-art effectiveness, surpassing strong baselines through synergistic integration of textual, facial, and visual modalities. The ablation study confirmed the necessity of each component, particularly attention-based fusion and temporal modeling, to achieve robust and generalizable results. Finally, the interpretability analysis via attention weights highlighted the dominant role of textual cues while validating the complementary contributions of nonverbal modalities. Collectively, these findings validate the design choices of the architecture and

emphasize its practical relevance for real-world, clinically informed applications. By achieving both high predictive accuracy and transparent interpretability, the proposed framework addresses the dual requirements of performance and trustworthiness in healthcare-oriented artificial intelligence.

Chapter 7

Conclusion and Future Direction

This chapter synthesizes the research presented in this thesis and reflects on its implications for the detection of depression using multimodal data. Building on the motivation and problem statement introduced in Chapters 1 and Chapters 2 and the framework detailed in Chapters 3 and Chapters 4, we evaluated the proposed Modality-Level Graph Attention (MLGA) architecture through a rigorous experimental program in Chapters 5 and Chapters 6. The overarching aim was to design and empirically evaluate an end-to-end system that integrates textual, visual, and facial cues, learns cross-modal relationships, models temporal dynamics, and is robust to noise and missing modalities.

We begin with a concise summary of the thesis contributions (Section 7.1), followed by a discussion of the research limitations (Section 7.3) that contextualize the scope and boundary conditions of our findings. Finally, we outline concrete directions for future work (Section 7.4), highlighting opportunities to extend the methodology, broaden its clinical applicability, and support responsible deployment.

7.1 Summary of Contributions

This thesis presents a comprehensive investigation into the use of deep learning for multimodal depression detection, addressing a critical gap at the intersection of artificial intelligence, psychology, and healthcare. This study makes several key contributions spanning methodological innovations, empirical findings, technical implementation, and interpretability, which are summarized below.

- **Development of a Novel Multimodal Framework:** A multimodal architecture was designed and implemented that integrated textual features from ClinicalBERT, visual features from VGG-PCA, and facial behavioral embeddings from OpenFace, which were fused using cross-modal attention and graph-based learning. By leveraging Graph Attention Networks (GAT) and a temporal sequence module, the framework successfully models inter-modality relationships and temporal dependencies, thereby advancing beyond conventional early- or late-fusion strategies.

- **Introduction of Modality-Level Graph Attention:** Unlike prior studies that operate pri-

marily at feature-level concatenation, this study introduced a modality-level GAT, where each modality is treated as a node in a participant-specific graph. This approach enables the model to dynamically weigh the contributions of the language, visual, and facial inputs, thereby improving adaptability and robustness. This represents a methodological contribution to the growing body of multimodal graph learning in affective computing.

- **Rigorous Evaluation Across Multiple Benchmark Datasets:** The framework was systematically validated using three benchmark corpora:
  - **E-DAIC (Distress Analysis Interview Corpus)** [29] for multimodal clinical interviews,
  - **EATD-Corpus** [30] for audio-textual depression detection, and
  - **D-Vlog** [31] for vlog-based depression analysis.

  By adopting the official dataset splits and evaluation metrics, this study ensured reproducibility and comparability with state-of-the-art baselines. The results demonstrated consistent improvements in the F1-score, ROC-AUC, and robustness across the corpora within these benchmarks, suggesting that the proposed approach is competitive across heterogeneous datasets under the studied conditions.

- **Addressing Class Imbalance and Regularisation:** This study incorporated Gaussian noise augmentation, modality dropout, and class-weighted loss functions to mitigate the issues of overfitting and class imbalance inherent in depression datasets. These strategies enhanced the model's sensitivity to minority (depressed) cases while maintaining high overall accuracy, thereby aligning the work with clinical priorities, where false negatives carry significant costs.

- **Advancements in Explainability:** To bridge the gap between performance and clinical trust, the architecture integrates modality-level attention weight visualization and graph-based interpretability methods. This represents an important step toward making deep learning models more transparent for use in mental health contexts, where explainability is crucial for clinician acceptance and regulatory approvals.

- **Empirical Benchmarking and Comparative Analysis:** Through the implementation of multiple baselines, including traditional classifiers (SVM, MLP), unimodal BERT-based systems, and late-fusion frameworks, this study provides a rigorous comparative analysis. The findings highlight the superiority of graph-based multimodal modelling over conventional fusion methods and establish new performance benchmarks for the E-DAIC, EATD-Corpus, and D-Vlog.

- **Practical Implementation and Lightweight Design:** A complete end-to-end pipeline was developed in **Google Colab**, enabling reproducible experimentation under real-world computing constraints. Importantly, the architecture was carefully designed with dimensionality reduction (PCA), parameter-efficient ClinicalBERT embeddings, and a modality-level GAT to maintain strong predictive performance while remaining computationally tractable. This balance indicates that, unlike many transformer-heavy multimodal frameworks, the proposed model may be adapted for deployment on mobile or edge devices with additional optimisation, validation, and systems engineering, opening pathways for real-world telehealth and digital mental health applications. Such transparent and efficient implementation not only supports academic reproducibility but also lowers the barriers for future researchers entering this interdisciplinary field.

- **Scholarly Contribution to the Field of Affective Computing:**

  Finally, this thesis consolidates insights from psychology, computational linguistics, and deep learning to advance the broader understanding of depression detection. By systematically articulating the limitations and mapping future research pathways, this study contributes to both academic knowledge and the translational potential of AI in clinical practice.

In conclusion, this study advances the state of the art in multimodal depression detection by developing a robust, interpretable, and empirically validated framework within the evaluated settings that is sensitive to the complexities of human behavior. These contributions provide a strong foundation and reference point for future interdisciplinary research aimed at building trustworthy, scalable, and clinically impactful AI systems for mental healthcare.

7.2 Research Questions and Their Answers

The experiments and analyses presented in Chapters 5 and 6 allow the research questions formulated in Section 1.4 to be revisited.

RQ1: Does a modality-level graph attention mechanism improve depression-detection performance compared with unimodal models and simple feature concatenation or late–fusion baselines on the E-DAIC-WOZ benchmark? The results on E-DAIC-WOZ show that the proposed modality-level GAT architecture achieves higher F1-scores and ROC–AUC than unimodal text-only, visual-only, or facial-only models, and than baselines based on simple feature concatenation and late fusion (see Tables 6.2 and 6.5). Removing the GAT layer and replacing it with concatenation (Model B in the ablation study) leads to a clear degradation in

both metrics, confirming that explicit graph-based attention over modalities yields a measurable performance gain. Thus, RQ1 is answered positively: modality-level graph attention improves depression-detection performance on E-DAIC relative to the considered baselines.

RQ2: What are the relative contributions of textual, visual, and facial modalities, and of temporal modelling and regularisation components, to the overall performance of the proposed framework? The ablation results in Table 6.5 indicate that removing any single modality (text, visual, or facial) reduces performance, with the largest drop occurring when the textual modality (ClinicalBERT) is removed, highlighting its dominant predictive role. Nonetheless, the exclusion of visual or facial inputs also leads to noticeable declines in F1-score and ROC–AUC, demonstrating that nonverbal cues provide complementary information. Likewise, removing the GRU temporal layer or the regularisation components (normalization, Gaussian noise, modality dropout) results in degraded generalisation. The attention-weight analysis in Table 6.6 further shows that text tends to receive the highest average attention, followed by visual and facial modalities. Collectively, these findings answer RQ2 by showing that all modalities and architectural components contribute, with text being most influential but visual and facial information, temporal modelling, and regularisation being necessary for the strongest and most stable performance.

RQ3: How robust is the proposed modality-level graph attention architecture under dataset and modality shifts, as assessed by transferring it from E-DAIC-WOZ to the EATD-Corpus and D-Vlog benchmarks? The cross-dataset evaluation in Chapter 6 reveals a mixed picture. On the one hand, when transferred without task-specific fine-tuning to the EATD-Corpus, and used in a text-only configuration under a clear language mismatch between Mandarin transcripts and an English-domain ClinicalBERT encoder, the model retains reasonable discriminatory power, particularly in terms of recall (Table 6.3). On the other hand, performance on the D-Vlog dataset, where only visual information is available in an in-the-wild setting, is substantially lower, with a ROC–AUC of 0.5885 and reduced F1-score (Tables 6.1 and 6.4). These findings indicate that the architecture exhibits some robustness under moderate domain and modality shifts, but that its out-of-domain performance, especially on uncontrolled visual-only data, is limited without targeted domain adaptation. RQ3 is therefore answered partially: the model generalizes reasonably under certain shifts, but strong real-world deployment in social-media-like settings would require additional adaptation and modelling work.

7.3 Limitations

Despite the advances achieved in this thesis, several limitations remain that must be considered when interpreting the results and assessing the generalizability of the proposed framework. These highlight the ongoing challenges in multimodal affective computing and point to future research directions. Three key limitations are emphasized below.

- **Demographic and Cultural Diversity of Datasets:** The datasets employed for both development and evaluation, namely E-DAIC [29, 106] and EATD-Corpus [30], were drawn from relatively homogeneous populations. E-DAIC participants were predominantly Western, whereas the EATD-Corpus consisted mainly of Mandarin-speaking university students. As cultural and linguistic contexts critically shape the manifestation of depressive symptoms, the model may not generalize effectively to more diverse groups [20]. Without broader cross-cultural validation, its utility in global telehealth and multicultural clinical practice remains uncertain.

- **Computational Complexity and Real-Time Constraints:** The architecture integrates resource-intensive components, including ClinicalBERT for text encoding, VGG-16 for visual representation, and Graph Attention Networks for fusion. Although dimensionality reduction (for example, PCA) and parameter-efficient embeddings mitigate some costs, the system remains demanding in terms of memory and inference time. Moreover, the analysis of model complexity in Chapter 5 focuses on the fusion module operating on precomputed embeddings and does not include the full ClinicalBERT and VGG-16 backbones, which remain computationally heavy. VGG-16 in particular is an older and comparatively large visual backbone; more recent, lightweight architectures could further reduce the computational budget. As a result, the current implementation is better viewed as a reference architecture under precomputed-embedding conditions than as a fully optimised mobile deployment. This constrains its suitability for mobile or low-resource clinical settings [145]. Further optimization strategies, such as pruning, quantization, knowledge distillation, or the adoption of more efficient encoders are needed for large-scale clinical deployment.

- **Limited Model Interpretability:** Although modality-level attention visualization was introduced, interpretability remains limited. Current outputs indicate only the relative contribution of modalities without fine-grained feature attribution or counterfactual reasoning. In sensitive domains, such as mental health, clinicians expect explainable AI systems that can provide transparent justifications for predictions [33, 146]. The absence of more advanced interpretability techniques constrains the clinical trustworthiness and regulatory readiness of the framework.

7.4 Scope for Future Work

Although this thesis achieved promising results, there are several directions in which the work can be extended to make the system more useful and reliable in practice. Three important areas for future research are highlighted.

- **Longitudinal and Temporal Modeling:** The current study focused on predicting a single depression label for each participant based on their interview session. A valuable next step would be to extend this approach to longitudinal analysis, in which depression levels can be tracked over weeks or months. Such models could support relapse monitoring, evaluate treatment progress, and provide early warnings before the symptoms worsen. Advances in sequential models, such as recurrent networks or transformers, can be explored to capture long-term changes more effectively.
- **Domain Adaptation and Cross-Cultural Generalization:** The datasets used in this thesis come from specific populations, which may limit how well the model works across different cultural and linguistic groups. Therefore, future work should focus on domain adaptation and transfer learning so that the system can adapt to new languages, cultural expressions, and recording environments. This would make the framework fairer, more inclusive, and suitable for global use in digital mental health.
- **Lightweight and Deployable Architectures:** While the proposed framework is more compact than many existing multimodal models, it still requires significant computational resources. For the system to be practical in telehealth or mobile applications, future studies should investigate methods such as pruning, quantization, or knowledge distillation to further reduce complexity. Creating a lighter version of the model would make real-time deployment possible on devices such as smartphones or tablets, increasing accessibility for communities with limited resources.

In summary, extending the framework to handle longitudinal data, ensuring cross-cultural generalization, and designing more efficient models for deployment are key steps that can help bridge the gap between research and real-world clinical application.

7.5 Vision for the Future

Looking ahead, the integration of advanced artificial intelligence with multimodal human behavioral data has the potential to fundamentally reshape the landscape of mental health care. By combining insights from language, facial expressions, vocal signals, physiological measures, and contextual information, future systems may evolve into truly personalized, continuous, and

culturally sensitive screening tools. Such systems would enable not only early detection but also proactive intervention and long-term monitoring of mental well-being of users.

The convergence of lightweight edge AI, privacy-preserving federated learning, and inherently explainable models promises to unlock new possibilities for remote telehealth, digital therapeutics, and unobtrusive, real-time monitoring in everyday environments. These advances could facilitate scalable and equitable access to care, particularly in underserved or resource-constrained communities where traditional clinical support is limited.

Ultimately, the emergence of intelligent, accessible, and trustworthy AI-driven platforms has the potential to bridge the global mental health gap. By empowering clinicians with interpretable decision support, equipping individuals with personalized feedback, and enriching our scientific understanding of depression as a dynamic and multifaceted condition, these systems can contribute meaningfully to clinical practice and public health.

However, realizing this vision will require continued interdisciplinary collaboration. AI researchers, clinicians, ethicists, and lived-experience communities must work together to ensure that innovations are not only technically robust, but also ethically grounded, culturally inclusive, and aligned with real-world needs and values.

7.6 Final Remarks

This thesis demonstrates that multimodal deep learning can provide a powerful and interpretable framework for automated depression detection, unifying linguistic, visual, and facial behavioral markers into a clinically meaningful system. By innovating through modality-level attention, temporal modeling, regularization strategies, and lightweight architectural choices, this study advances both the theoretical understanding and practical applicability of the proposed approach. Empirical evaluations across three benchmark datasets (E-DAIC, EATD-Corpus, and D-Vlog) confirmed the effectiveness of the proposed framework on E-DAIC and showed competitive performance under zero-shot and visual-only settings on EATD-Corpus and D-Vlog, while also revealing clear limitations under language and domain shift.

While limitations remain, the contributions outlined in this study provide a strong foundation for continued interdisciplinary progress. Importantly, this study addressed the central research question of whether multimodal fusion, when modeled through graph attention and temporal dynamics, can improve the accuracy and interpretability of depression detection. The findings affirm this hypothesis and highlight the advantages of modality-level graph-based approaches.

Looking ahead, the convergence of technical innovation with clinical needs highlights a profound opportunity: the development of accessible, trustworthy, and culturally sensitive AI systems that can support mental healthcare worldwide. By expanding modalities, enhancing interpretability, and optimizing deployment for mobile and telehealth environments, future

research can translate this vision into practice in the near future.

Ultimately, this thesis contributes not only a novel model but also a pathway toward more empathetic, personalized, and inclusive digital mental health solutions, which is an important step toward narrowing the global mental health gap. Taken together, this study underscores the promise of AI not as a replacement for clinical expertise, but as a vital augmentation that can extend care, enable early detection, and strengthen mental health support globally.

Bibliography

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *In Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

[2] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical bert embeddings," *In Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 72–78, 2019.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations (ICLR)*, 2015.

[4] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[5] World Health Organization, "Depression," *World Health Organization Fact Sheets*, 2023, accessed: 2025-08-20. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression

[6] T. Vos, S. S. Lim, C. Abbafati, and et al., "Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019," *The Lancet*, vol. 396, no. 10258, pp. 1204–1222, 2020.

[7] A. J. Ferrari, F. J. Charlson, R. E. Norman, and et al., "Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010," *PLoS Medicine*, vol. 10, no. 11, p. e1001547, 2013.

[8] K. Kroenke, R. L. Spitzer, and J. B. Williams, "The phq-9: validity of a brief depression severity measure," *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 606–613, 2001.

[9] A. T. Beck, R. A. Steer, and G. K. Brown, "Beck depression inventory-ii," *San Antonio*, vol. 78, no. 2, pp. 490–498, 1996.

[10] V. Patel, S. Xiao, H. Chen, and et al., "The magnitude of and health system responses to the mental health treatment gap in adults in india and china," *The Lancet*, vol. 388, no. 10063, pp. 3074–3084, 2016.

[11] N. Ramirez-Esparza, C. K. Chung, E. Kacewicz, and J. W. Pennebaker, "Language as a marker of depression: Perspectives from computational linguistics," *Current Opinion in Psychology*, vol. 4, pp. 11–16, 2018.

[12] S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, and H. Ohsaki, "Detecting depression in twitter using real-time mental health surveys," *In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 3187–3196, 2015.

[13] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre, "Detecting depression from facial actions and vocal prosody," *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pp. 1–7, 2009.

[14] H. Dibeklioglu, Z. Hammal, Y. Yang, and J. F. Cohn, "Multimodal depression detection: A systematic review," *Journal on Multimodal User Interfaces*, vol. 12, no. 2, pp. 143–158, 2018.

[15] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[16] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[17] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial, verbal, and vocal cues," in *IEEE Transactions on Affective Computing*, vol. 10, no. 1. IEEE, 2018, pp. 3–16.

[18] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 96–108, 2017.

[19] Y. Yang, X. Li, and B. Sun, "Deep multimodal fusion by channel exchanging," *Information Fusion*, vol. 65, pp. 54–64, 2021.

[20] B. Inkster, S. Sarda, and V. Subramanian, "Machine learning and mental health: Possibilities and pitfalls," *Trends in Cognitive Sciences*, vol. 22, no. 4, pp. 291–293, 2018.

[21] R. A. Calvo, D. N. Milne, M. S. Hussain, and H. Christensen, "Natural language processing in mental health applications using non-clinical texts," *Natural Language Engineering*, vol. 23, no. 5, pp. 649–685, 2017.

[22] J. Torous, K. Jän Myrick, N. RauseoRicupero, and J. Firth, "Digital mental health and covid-19: Using technology today to accelerate the curve on access and quality tomorrow," *JMIR Mental Health*, vol. 7, no. 3, p. e18848, 2020.

[23] P. Cuijpers, E. Karyotaki, E. Weitz, G. Andersson, S. D. Hollon, and A. van Straten, "Psychotherapy for depression in adults: a meta-analysis of comparative outcome studies," *Journal of Consulting and Clinical Psychology*, vol. 82, no. 6, pp. 970–986, 2014.

[24] S. Evans-Lacko and M. Knapp, "Global patterns of workplace productivity for people with depression: absenteeism and presenteeism costs across eight diverse countries," *Social Psychiatry and Psychiatric Epidemiology*, vol. 51, pp. 1525–1537, 2016. [Online]. Available: https://link.springer.com/article/10.1007/s00127-016-1278-4

[25] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, D. P. Rosenwald, and S. Wilkinson, "Social risk and depression: Evidence from facial expressions," *Psychological Science*, vol. 25, no. 11, pp. 2117–2128, 2014.

[26] T. Alhanai, M. Ghassemi, and J. Glass, "Detecting depression with audio/text sequence modeling of interviews," *Interspeech 2018*, pp. 1716–1720, 2018.

[27] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, 2021.

[28] X. Gao, B. Sun, Q. Yin, and D. Meng, "Modality dropout for multimodal emotion recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 1176–1187, 2021.

[29] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella *et al.*, "The distress analysis interview corpus of human and computer interviews," in *Proceedings of LREC*, 2014, pp. 3123–3128.

[30] B. Shen, L. Zhang, S. Liu, H. Xue, X. Wang, Z. Xu, and M. Zhao, "The eatd-corpus: Emotional audio-textual depression corpus for depression detection in mandarin," in *Interspeech 2022*, 2022, pp. 2933–2937.

[31] J. Yoon, C. Kang, S. Kim, and J. Han, "D-vlog: Multimodal vlog dataset for depression detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[32] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *International Conference on Learning Representations (ICLR)*, 2018.

[33] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[34] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[35] A. B. Shatte, D. M. Hutchinson, and S. J. Teague, "Machine learning in mental health: A scoping review of methods and applications," *Psychological Medicine*, vol. 49, no. 9, pp. 1426–1448, 2019.

[36] M. Liu, M. Zhang, F. Wu, and X. Xie, "A survey on lightweight deep learning models for resource-constrained devices," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[37] N. D. Lane, S. Bhattacharya, P. Georgiev, C. Forlivesi, F. Kawsar, and A. R. Beresford, "Can deep learning revolutionize mobile sensing?" *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*, pp. 117–122, 2015.

[38] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable ai systems for the medical domain?" *arXiv preprint arXiv:1712.09923*, 2017.

[39] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, 2021.

[40] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes *et al.*, "Model cards for model reporting," in *FAT**, 2019.

[41] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan *et al.*, "Datasheets for datasets," *Communications of the ACM*, 2021.

[42] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in twitter," *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (CLPsych)*, pp. 51–60, 2014.

[43] P. Resnik, W. Armstrong, L. Claudino, V.-A. Nguyen, T. Nguyen, and J. Boyd-Graber, "Beyond lda: Exploring supervised topic modeling for depression-related language in twitter," *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pp. 99–107, 2015.

[44] W. Yin, K. Kann, M. Yu, and H. Schütze, "A comparative study of deep learning models for sentiment analysis," *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2766–2777, 2017.

[45] K. Huang, J. Altosaar, and R. Ranganath, "Clinical nlp: Progress and challenges," *Annual Review of Biomedical Data Science*, vol. 5, pp. 363–390, 2022.

[46] Z. Wang, H. He, Y. Chen, and J. Wang, "Multimodal depression detection: A review of trends, challenges, and opportunities," *IEEE Transactions on Affective Computing*, 2022.

[47] S. Chancellor and M. De Choudhury, "Methods in predictive techniques for mental health status on social media: A critical review," *npj Digital Medicine*, vol. 3, no. 1, pp. 1–11, 2020.

[48] H. L. Kim and E. L. Murnane, "Social media and depression: Can machine learning tools help?" *Journal of the American Medical Informatics Association*, vol. 27, no. 12, pp. 1941–1947, 2020.

[49] M. S. R. Al Hossain, M. T. Afzal, and U. Farooq, "Detecting depression in arabic tweets using machine learning," *Cognitive Systems Research*, vol. 54, pp. 482–494, 2018.

[50] S. Amiriparian, M. Gerczuk, and B. W. Schuller, "Snore sound classification using image-based deep spectrum features," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 3512–3516.

[51] X. Liu, R. Wang, R. Li, and K. Wang, "Multi-view multimodal fusion for depression detection," *Sensors*, vol. 21, no. 24, p. 8436, 2021.

[52] F. Ringeval, B. W. Schuller, M. Valstar *et al.*, "Avec 2017: Real-life depression, and affect recognition workshop and challenge," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. Association for Computing Machinery, 2017, pp. 3–9.

[53] P. Dham, Y. Wang, A. Singh, R. Bhatia, and S. Kumar, "Multimodal depression detection: A systematic review and meta-analysis," *IEEE Transactions on Affective Computing*, 2023, early Access.

[54] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: An open source facial behavior analysis toolkit," *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10, 2016.

[55] T. Tang, Y. Han, J. Li, and S. Yu, "Vit-based multimodal depression detection using facial landmarks and speech," *IEEE Transactions on Affective Computing*, 2023.

[56] D. Zhu, J. Yin, X. Wang, X. Ren, and R. Wang, "Multi-scale and cross-modal fusion for video-based depression recognition," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4103–4107.

[57] Y. Yang, Y. Li, X. Huang, and F. Chen, "Depression recognition from multimodal information with temporal fusion," *Neural Computing and Applications*, vol. 35, pp. 15 997–16 013, 2023.

[58] M. R. Morales, S. I. Levitan, and S. Scherer, "A crossmodal review of multimodal depression detection," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2018, pp. 1–8.

[59] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, R. Salakhutdinov, and L.-P. Morency, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019, pp. 6558–6569.

[60] S. Poria, E. Cambria, and A. Gelbukh, "Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research," *IEEE Transactions on Affective Computing*, vol. 12, no. 3, pp. 383–393, 2020.

[61] A. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multiview sequential learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[62] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[63] P. P. Liang, T. Li, A. Zadeh, and L.-P. Morency, "Interpretable multimodal depression detection in social media: A case study and user perspective," *IEEE Transactions on Affective Computing*, 2023.

[64] W. N. Price and I. G. Cohen, "Ethics and governance of artificial intelligence for health," *World Health Organization*, 2022. [Online]. Available: https://www.who.int/publications/i/item/9789240029200

[65] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations (ICLR)*, 2015.

[66] Q. Zhang, Y. Yang, H. Wang, Z. Jin, and Y. Zhou, "A survey on attention mechanisms in deep learning," *Neurocomputing*, vol. 489, pp. 126–148, 2022.

[67] J. Zhao, Z. Zhang, Y. Xu, and M. Zhou, "Transformer-based multimodal depression detection with dynamic modality interaction," *Information Fusion*, vol. 86, pp. 84–95, 2022.

[68] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating long sequences with sparse transformers," *arXiv preprint arXiv:1904.10509*, 2019.

[69] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, "Linformer: Self-attention with linear complexity," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 5516–5526.

[70] K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser *et al.*, "Rethinking attention with performers," in *International Conference on Learning Representations*, 2021.

[71] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," in *EMNLP-IJCNLP*, 2019.

[72] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *NeurIPS*, 2019.

[73] Y.-C. Chen *et al.*, "Uniter: Universal image-text representation learning," in *ECCV*, 2020.

[74] H. Liu, L. Wang, and Y. Wu, "Efficient multimodal attention fusion for mobile affective computing," *IEEE Transactions on Affective Computing*, 2022.

[75] L. Chen, Y. Wang, H. Wang, and B. Hu, "Joint cross-modal attention for multimodal depression detection," in *Proceedings of the 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2023, pp. 1423–1426.

[76] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Advances in Neural Information Processing Systems*, 2016, pp. 289–297.

[77] Y. Li, X. Zheng, Y. Liu, W. Wang, Z. Yang, and T. Wang, "Hierarchical attention networks for document classification," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2018, pp. 1480–1489.

[78] M. Xu, C. Xu, and T. M. Hospedales, "Hierarchical multimodal transformer for efficient long video understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[79] Y. Xia, L. Liu, T. Dong, J. Chen, Y. Cheng, and L. Tang, "Depression detection model based on multimodal graph neural network," *IEEE Transactions on Affective Computing*, 2023.

[80] J. Rieger, J. Timmermann, C. Sassenrath, H. Nisar, S. Yigit, N. Memon, and A. Reuter, "Interpretability and explainability in ai: A review," *IEEE Access*, vol. 8, pp. 206 299–206 316, 2020.

[81] S. Jain and B. C. Wallace, "Attention is not explanation," 2019.

[82] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019, pp. 11–20.

[83] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," *ACM Computing Surveys*, 2020.

[84] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: Grids, groups, graphs, geodesics, and gauges," *arXiv preprint arXiv:2104.13478*, 2021. [Online]. Available: https://arxiv.org/abs/2104.13478

[85] X. Li, Y. Dong, Y. Yi, Z. Liang, and S. Yan, "Hypergraph neural network for multimodal depression recognition," *Electronics*, vol. 13, no. 22, p. 4544, 2024. [Online]. Available: https://www.mdpi.com/2079-9292/13/22/4544

[86] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *ICLR*, 2017.

[87] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *NeurIPS*, 2017.

[88] E. Rossi, B. Chamberlain, F. Frasca, D. Eynard, M. M. Bronstein, and F. Monti, "Temporal graph networks for deep learning on dynamic graphs," *arXiv:2006.10637*, 2020.

[89] D. Xu *et al.*, "Inductive representation learning on temporal graphs," in *ICLR*, 2020.

[90] A. Sankar *et al.*, "Dysat: Deep neural representation learning on dynamic graphs via self-attention networks," in *WSDM*, 2020.

[91] R. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "Gnnexplainer: Generating explanations for graph neural networks," in *NeurIPS*, 2019.

[92] Y. Shen, W. Liu, Y. Li, Y. Wu, and J. Wang, "Multimodal graph fusion for depression detection in social media," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 2290–2299.

[93] W. Hu, B. Liu, J. Gomes, M. Zitnik, P. Liang, V. Pande, and J. Leskovec, "Strategies for pre-training graph neural networks," in *ICLR*, 2020.

[94] C. Ying, T. Cai, S. Luo, S. Zheng, G. Ke, D. He, L. Shen, and T.-Y. Liu, "Do transformers really perform badly on graphs?" in *NeurIPS*, 2021.

[95] T. Baltrušaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. Xi'an, China: IEEE, May 2018, pp. 59–66. [Online]. Available: https://multicomp.cs.cmu.edu/wp-content/uploads/2018/11/OpenFace.pdf

[96] C. M. Bishop, "Training with noise is equivalent to tikhonov regularization," in *Neural Computation*, vol. 7, no. 1. MIT Press, 1995, pp. 108–116.

[97] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.

[98] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, p. e0118432, 2015.

[99] I. T. Jolliffe and J. Cadima, *Principal Component Analysis*, 3rd ed. Springer, 2021.

[100] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto, CA: Consulting Psychologists Press, 1978.

[101] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[102] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: https://www.deeplearningbook.org/

[103] M. Shukla, A. Rajput, M. Kumari, and P. Kumaraguru, "Multimodal graph neural networks for automatic depression recognition," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 2974–2982.

[104] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *International Conference on Machine Learning*, pp. 448–456, 2015.

[105] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

[106] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt *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*. Nice, France: ACM, 2019, pp. 3–12.

[107] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, and D. R. Traum, "The distress analysis interview corpus of human and computer interviews," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 3123–3128.

[108] W. W. Zung, "A self-rating depression scale," *Archives of General Psychiatry*, vol. 12, no. 1, pp. 63–70, 1965.

[109] J. Guo, T. Shen, J. Liu, T. Sun, and J. Zhou, "D-vlog: In-the-wild benchmark for visual depression detection in vlogs," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1045–1057, 2023.

[110] H. Jiang, Y. Li, J. Guo, T. Shen, Y. Tang, H. Xu, and L. Zhang, "A survey on deep learning for depression detection: Datasets, methods, and challenges," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 834–849, 2023.

[111] M. Zhao, J. Wu, J. Luo, J. Jiang, and Q. M. Chen, "Depression detection from social media posts based on user behaviors and deep neural networks," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 5, pp. 1147–1159, 2021.

[112] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, "Psychological aspects of natural language use: Our words, our selves," *Annual Review of Psychology*, vol. 54, no. 1, pp. 547–577, 2003.

[113] S. S. Rude, E.-M. Gortner, and J. W. Pennebaker, "Language use of depressed and depression-vulnerable college students," *Cognition & Emotion*, vol. 18, no. 8, pp. 1121–1133, 2004.

[114] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific Data*, vol. 3, p. 160035, 2016.

[115] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[116] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "MentalBERT: Publicly available pretrained language models for mental healthcare," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, H. Mazo, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, 2022, pp. 7184–7190. [Online]. Available: https://aclanthology.org/2022.lrec-1.778/

[117] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Knight, A. Nenkova, and O. Rambow, Eds. San Diego, California: Association for Computational Linguistics, 2016, pp. 1480–1489. [Online]. Available: https://aclanthology.org/N16-1174/

[118] G. Bradski, "The opencv library," *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 2000.

[119] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[120] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019.

[121] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 212–220.

[122] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations (ICLR)*, 2018, arXiv preprint arXiv:1710.09412.

[123] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[124] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, vol. 2. Morgan Kaufmann, 1995, pp. 1137–1143.

[125] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[126] A. Paszke, S. Gross, F. Massa *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *NeurIPS*, 2019.

[127] T. Wolf, L. Debut, V. Sanh *et al.*, "Transformers: State-of-the-art natural language processing," in *EMNLP: System Demonstrations*, 2020, pp. 38–45.

[128] M. Fey and J. E. Lenssen, "Fast graph representation learning with pytorch geometric," *ICLR Workshop*, 2019.

[129] M. Wang, L. Yu, D. Zheng, Q. Gan, Y. Gai, Z. Ye, M. Li, J. Zhou, Q. Huang, C. Ma, Z. Huang, Q. Guo, H. Zhang, H. Lin, J. Wu, C. Song, Z. Xiang, Z. Zhu, J. Ma, Z. Yang, L. Zhou, and G. K. Li, "Deep graph library: A graph-centric, highly-performant package for graph neural networks," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '19)*. ACM, 2019, pp. 2201–2204.

[130] C. R. Harris *et al.*, "Array programming with numpy," *Nature*, vol. 585, pp. 357–362.

[131] W. McKinney, "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, 2010.

[132] F. Pedregosa, G. Varoquaux, A. Gramfort *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[133] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," in *arXiv preprint arXiv:1704.04861*, 2017.

[134] L. M. Huynh, Y. Tan, and Y. W. Lee, "Edge ai: On-demand accelerating deep neural network inference via edge computing," *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 447–457, 2020.

[135] K. Denecke, A. V. Cimino, C. A. Madani, and N. Abidi, "Explainable artificial intelligence for mental health through multidimensional classification and visualization," *Frontiers in Psychiatry*, vol. 12, p. 646501, 2021.

[136] A. J. Mitchell, "Clinical significance of questionnaire measurement of depression: A meta-review of diagnostic validity," *British Journal of Psychiatry*, vol. 194, no. 4, pp. 321–328, 2009.

[137] A. Sarda, A. Begum, J. Zhou, and X. Li, "Mental health detection using machine learning: A review," *Current Psychiatry Reports*, vol. 24, no. 5, pp. 205–220, 2022.

[138] J. Jung, C. Kang, J. Yoon, S. Kim, and J. Han, "Hique: Hierarchical question embedding network for multimodal depression detection," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24)*. Boise, ID, USA: Association for Computing Machinery, 2024, pp. 1049–1059. [Online]. Available: https://arxiv.org/abs/2408.03648

[139] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 4190–4197.

[140] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[141] Z. Huang, Z. Ren, M. Dong, and B. W. Schuller, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. INTERSPEECH*, 2020, pp. 232–236.

[142] Y. Xia, L. Liu, T. Dong, J. Chen, Y. Cheng, and L. Tang, "A hybrid bert-cnn approach for depression detection on social media using multimodal data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[143] T. Shen, J. Wang, T. Sun, J. Liu, X. Zhu, and J. Zhou, "Multimodal fusion with graph neural networks for depression detection," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2022, pp. 4348–4356.

[144] E. Clark, A. Radford, F. de la Torre, and L.-P. Morency, "Interpretable multimodal depression prediction with attention-based fusion," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1–13, 2020.

[145] S. Han, H. Mao, and W. J. Dally, “Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding,” in *International Conference on Learning Representations (ICLR)*, 2016, conference version published in 2016; originally posted to arXiv in 2015. [Online]. Available: https://arxiv.org/abs/1510.00149

[146] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. [Online]. Available: https://arxiv.org/abs/1705.07874