Beyond Text: Multimodal Analysis for Mental Health Diagnostics leveraging Large Language Models

A THESIS

submitted by

CHAYAN TANK

(MT23030)

in partial fulfilment of the requirements

for the award of the degree of

MASTER OF TECHNOLOGY

COMPUTER SCIENCE AND ENGINEERING

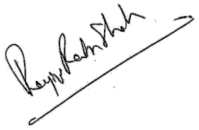
INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI

NEW DELHI- 110020

July 2025

THESIS CERTIFICATE

This is to certify that the thesis titled **Beyond Text: Multimodal Analysis for Mental Health Diagnostics leveraging Large language models**, submitted by **Chayan Tank**, to the Indraprastha Institute of Information Technology, Delhi, for the award of the degree of **M.Tech**, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other institution or University for the award of any degree or diploma.



Dr. Rajiv Ratn Shah
Thesis Supervisor
Associate Professor, Institute Chair Professor
Department of CSE and HCD
IIIT Delhi, 110020

Place: New Delhi

Date: July 2025

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to everyone who has supported me throughout the journey of this thesis.

First and foremost, I am profoundly thankful to my advisor, Dr. Rajiv Ratn Shah, for their invaluable guidance, encouragement, and unwavering support. Their deep insight into both machine learning and mental health domains greatly shaped the direction and quality of this work.

I would also like to thank the research students of the MIDAS Lab for their guidance, collaboration, feedback, and camaraderie during the development of this work. I am particularly grateful to Dr. Avinash Anand, PhD scholar from the research lab, for their exceptional mentoring, constant support, and guidance throughout my work. Without their invaluable guidance, this thesis would not have been possible.

I am deeply thankful to my friends Sarthak Pol, Vinayak Katoch, Shaina Mehta, and Sonik Sarungale for their constant support and assistance throughout the thesis and project work. Their help and encouragement were invaluable during challenging times.

This research would not have been possible without the invaluable contributions of the open-source community and mental health professionals. I gratefully acknowledge the developers of open-source LLMs, whose innovations form the backbone of modern language understanding, as well as the mental health experts whose work continues to inspire better integration of AI in socially impactful domains.

Lastly, I would like to express my deepest gratitude to my parents and siblings for their unwavering support, patience, and belief in me. Their constant encouragement and understanding were vital during challenging times and milestones alike.

This thesis is dedicated to all those who work tirelessly to improve mental health systems and to those who struggle silently; may technology serve as a bridge to support, care, and understanding.

ABSTRACT

KEYWORDS: Mental Health Assessment ; Large Language Models ; Multimodality ; Depression and Suicide Risk Prediction

Mental health disorders, particularly depression and suicide risk, represent critical public health challenges that are often underdiagnosed due to social stigma, lack of access to professionals, and the limitations of traditional diagnostic tools such as questionnaires. This thesis consolidates a series of research contributions that explore the use of Large Language Models (LLMs), both unimodal and multimodal, for more scalable, accurate, and context-aware mental health assessment.

The analysis encompasses three key approaches of analysis: Establishing baseline performance using classical ML/DL models on the E-DAIC and Reddit datasets with traditional feature extraction and fusion techniques; comprehensive benchmarking of state-of-the-art LLMs in zero-shot and few-shot settings for depression and suicide risk prediction; and evaluating the EDAIC test set on the novel AM-LLM framework which introduces a model-agnostic, multilingual architecture that combines audio and text for enhanced mental health assessment, demonstrating improved performance in depression detection tasks than just textual analysis, with comparable performance on English and Hindi languages.

The thesis also provides a critical perspective on the scalability, bias, and ethical implications of deploying LLMs in sensitive health contexts and explores building explainable mental health support systems. Collectively, this work demonstrates that multimodal LLMs, when properly adapted and evaluated, hold immense promise for augmenting mental health diagnosis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AI	Artificial Intelligence
SOTA	State of the Art
NLP	Natural Language Processing
AM-LLM	Audio-Text Multimodal LLM
ASR	Automatic Speech Recognition
BiLSTM	Bidirectional LSTM
CCC	Concordance Correlation Coefficient
CNN	Convolutional Neural Network
E-DAIC	Extended Distress Analysis Interview Corpus
F1	F1-Score (harmonic mean of precision, recall)
GPT	Generative Pre-trained Transformer
LLM	Large Language Model
MAE	Mean Absolute Error
MFCC	Mel-frequency cepstral coefficients
OS-DVD	Open Source Depression Video Dataset
PHQ-8	Patient Health Questionnaire-8
RMSE	Root mean square error
RoBERTa	Robustly Optimized BERT Pretraining
SVM	Support Vector Machine
SUROBERTA	Suicide Risk RoBERTa
TTS	Text-to-Speech
WHO	World Health Organization

NOTATION

$\boldsymbol{N}$	Total number of samples
$\boldsymbol{K}$	Total number of classes
$\boldsymbol{y_i}$	True class label for sample i
$\boldsymbol{\hat{y}_i}$	Predicted class label or value for sample i
$\boldsymbol{a_i}$	Ground truth value for sample i (regression)
$\mathbf{1}(\boldsymbol{y_i} = \boldsymbol{\hat{y}_i})$	Indicator function (1 if prediction is correct, else 0)
$\boldsymbol{N_i}$	Number of true instances in class i
$\text{F1}_{\boldsymbol{i}}$	F1-score for class i
$\lvert\boldsymbol{a_i} - \boldsymbol{\hat{a}_i}\rvert$	Absolute error for sample i
$\boldsymbol{\eta}$	Pearson correlation coefficient
$\boldsymbol{\sigma_a}$	Standard deviation of ground truth values
$\boldsymbol{\sigma_b}$	Standard deviation of predicted values
$\boldsymbol{\alpha}$	Mean of ground truth values
$\boldsymbol{\beta}$	Mean of predicted values
F1	F1-score (harmonic mean of precision and recall)
wF1	Weighted F1-score
F1_{micro}	Micro-averaged F1-score (global precision/recall)
Accuracy	Proportion of correctly predicted instances
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
Cls	Classification task
Reg	Regression task

CHAPTER 1

INTRODUCTION

1.1 Overview

Mental health plays a crucial role in how people think, feel, and interact with the world. Today, conditions like depression and suicide risk have become major global health concerns. According to the World Health Organisation (WHO), nearly 1 in 8 people worldwide live with a mental health disorder, and over 700,000 people die by suicide each year, making it one of the leading causes of death among young people [1, 2].

Despite the seriousness of the problem, most people do not receive the help they need. In many low- and middle-income countries, up to 75% of people with mental health conditions receive no treatment, due to lack of access, stigma, or cost [1]. Even in places where care is available, traditional methods—like questionnaires and interviews-can be slow, subjective, and often miss early warning signs. There is also a case of reported bias [3]. The economic cost of untreated mental health issues is also massive. Depression and anxiety alone cost the global economy over $1 trillion per year in lost productivity [4].

At the same time, people are sharing more of their lives online through social media, messages, voice, and video. Research shows that changes in how people speak or post online can reveal early signs of depression or suicidal thoughts, often before they seek help [5, 6]. This has opened up new opportunities for using AI tools to detect mental health risks in a faster, more objective, and more scalable way.

In particular, Large Language Models (LLMs) and other AI systems have shown promise in analyzing text, speech, and even behavior patterns to support early detection, personalized insights, and mental health interventions. This thesis systematically explores how multimodal LLMs can be harnessed to create mental health diagnostic tools, with a focus on integrating audio and textual data for more accurate and equitable assessment across diverse populations.

1.2 Motivation and Core Hypothesis

While digital mental health tools have proliferated in recent years, most rely on unimodal data—primarily text—limiting their ability to capture the full spectrum of human communication, which includes speech, prosody, and nonverbal cues [7, 8]. Recent advances in large language models (LLMs) have demonstrated impressive capabilities in natural language understanding and generation, but their application to multimodal and multilingual mental health assessment remains underexplored [9, 10]. Notably, studies have shown that integrating audio and linguistic features can significantly improve the detection of affective states and mental health conditions, particularly in diverse and underrepresented populations [7, 9]. However, challenges persist in developing robust, generalizable models that can operate across languages and cultural contexts, as well as in establishing standardized benchmarks and evaluation protocols [11, 9].

This thesis is motivated by the need to bridge these gaps by systematically investigating the potential of multimodal LLMs for mental health assessment. The thesis hypothesizes that combining audio and text modalities in the LLM framework will yield more accurate and equitable detection of depression and related conditions. This work evaluates the effectiveness of recent advances in multimodal LLMs, including zero-shot performance and cross-lingual adaptation, for improving the scalability, accessibility, and cultural sensitivity of mental health assessment.

The limitations of traditional diagnostic tools have become increasingly apparent in our digital age. Subjectivity in assessment, resource intensity, and persistent social stigma continue to hinder effective diagnosis and timely intervention. However, the proliferation of smartphones and social media has created vast, real-world datasets containing rich linguistic and behavioral signals that can be leveraged for mental health assessment. Furthermore, recent advances in AI and Large Language Models (LLMs) now offer unprecedented capabilities to process, understand, and reason over complex, multi-source data—including text, audio, and video—at scale. The global nature of mental health challenges also demands tools that work across languages and cultures, motivating the development of multilingual and language-agnostic frameworks.

1.3 Research Evolution and Methodological Arc

This thesis presents a systematic exploration of mental health detection, tracing the evolution from classical machine learning to SOTALLMs. The research journey began with establishing baseline performance using classical ML models on the E-DAIC and Reddit suicide risk datasets. This initial phase involved experimentation with traditional features from audio and video data, including MFCCs and AU pairs, and evaluating models such as SVMs, BiLSTM, and modality fusion techniques using the OpenAI Whisper Encoder-decoder model for audio encodings [12]. These experiments revealed the limitations of unimodal, hand-crafted feature approaches, particularly in capturing the complex interplay between visual, linguistic, and acoustic signals.

Building on these baselines, the research progressed to leveraging the power of Large Language Models for mental health detection. This phase involved benchmarking the zero-shot and few-shot capabilities of models like GPT-4, LLaMA, and RoBERTa on depression and suicide risk tasks. The results demonstrated superior contextual understanding and generalization compared to classical models, while also highlighting the importance of prompt engineering and model scale in achieving optimal performance.

Recognizing the critical importance of suicide prevention, the research then focused on developing SUROBERTA, a semi-supervised, fine-tuned LLM for suicide risk prediction. This work addressed the challenges of class imbalance through data augmentation and pseudo-labeling, demonstrating that efficient, smaller models can achieve competitive results with minimal supervision. The approach showed particular promise in handling the complex, nuanced nature of suicide risk assessment from social media content.

The culmination of this research is the development of the AM-LLM framework, which directly tests the core hypothesis of multimodal integration. This framework combines audio and textual analysis in a unified, model-agnostic, multilingual architecture, demonstrating a 10% performance improvement over text-only analysis on depression detection tasks. The framework's multilingual capabilities were validated on both English and Hindi data, using the E-DAIC dataset.

1.4 Empirical Evidence and Real-World Impact

The experiments conducted in this research demonstrate the superiority of multimodal LLMs over unimodal and classical approaches, particularly in challenging, real-world data scenarios, without the need for extensive feature extraction and fine-tuning or training the models. The integration of audio and text modalities has proven especially valuable, with multimodal models achieving higher accuracy and robustness than their text-only counterparts. The multilingual frameworks developed in this work have shown promising generalization capabilities across languages, potentially reducing cultural and linguistic barriers to mental health assessment.

The use of real-world, diverse datasets, including E-DAIC and Reddit suicide datase,t has ensured the clinical relevance and scalability of the developed approaches. These datasets represent different aspects of mental health assessment, from clinical interviews to social media content, providing a comprehensive testbed for evaluating the robustness of the proposed methods. The research has also highlighted the critical role of prompt engineering and model selection in maximizing LLM performance in sensitive health domains.

1.5 Contributions of the Thesis

This thesis aims to make several contributions to the intersection of AI and mental health. The work establishes comprehensive baselines for classical ML approaches on E-DAIC and Reddit datasets, providing valuable reference points for future research. It presents a systematic assessment of LLMs for depression and suicide risk detection, including a detailed analysis of prompt engineering and model scale effects. The development of SUROBERTA introduces a novel, semi-supervised approach to suicide risk prediction, demonstrating efficient learning with minimal supervision.

The AM-LLM framework represents a major contribution in the form of a multilingual, model-agnostic audio-text framework, validated on the EDAIC dataset. The framework's 10% improvement over text-only analysis provides strong evidence for the value of multimodal integration in mental health assessment.

1.6 Structure of the Thesis

The remainder of this thesis is organized as follows:

- **Chapter 2: Literature Review** - Reviews related work, tracing the evolution from classical to LLM-based and multimodal approaches in mental health assessment.
- **Chapter 3: Classical Machine Learning and Neural Network Approaches** - Details baseline experiments using classical ML on E-DAIC and Reddit datasets, establishing fundamental performance metrics.
- **Chapter 4: LLM-based Analysis** - Presents LLM-based analysis for depression and suicide risk detection, exploring the capabilities and limitations of these models.
- **Chapter 5: AM-LLM Framework** - Introduces and validates the AM-LLM framework, including multilingual and multimodal experiments.
- **Chapter 6: Discussion** - Discusses broader implications, limitations, and future directions of the research.
- **Chapter 7: Conclusion** - Concludes with a summary of findings and their significance for mental health care.

CHAPTER 2

LITERATURE REVIEW

This chapter presents a comprehensive review of the literature on mental health diagnostics, focusing on the evolution from traditional methods to modern approaches that combine audio, visual, and textual analysis. We examine how the integration of these modalities has led to more accurate and reliable depression detection. Traditional assessments, such as self-reported questionnaires and clinical interviews, remain limited by subjectivity, accessibility barriers, and social stigma, often resulting in misdiagnosis. The field has since progressed through computational approaches, with early machine learning models leveraging hand-crafted features from audio, video, and text, and multimodal deep learning frameworks demonstrating the value of integrating multiple data sources. The advent of deep learning, particularly BiLSTM and attention-based architectures, enabled more effective modeling of temporal and semantic patterns in speech and language. Most recently, large language models (LLMs) and multimodal transformers have revolutionized the field, offering robust contextual understanding and the ability to process both linguistic and paralinguistic cues. Benchmarking studies show that these modern, multimodal architectures outperform classical methods, with improvements attributed to advanced feature fusion and cross-modal reasoning. Despite these advances, challenges remain in developing low-compute, scalable, clinically validated, and explainable tools that generalize across languages and populations.

2.1 Traditional Mental Health Assessment

Mental health disorders, particularly depression, anxiety, and suicide risk, remain pressing global health concerns, impacting hundreds of millions of individuals worldwide. According to the World Health Organization, more than 280 million people are affected by depression, and over 301 million individuals live with anxiety disorders globally [13]. The detection and diagnosis of these conditions have traditionally relied on standardized self-report questionnaires and structured clinical interviews.

Among the most widely used screening tools is the Patient Health Questionnaire-9 (PHQ-9), a brief 9-item scale designed to assess the severity of depressive symptoms and guide further diagnostic steps [14]. Similarly, the Generalized Anxiety Disorder 7-item scale (GAD-7) is frequently employed to evaluate generalized anxiety symptoms in primary care and clinical settings [15]. These tools are valued for their simplicity, ease of use, and validated psychometric properties, making them accessible for both clinical and research applications.

In addition to self-report scales, structured clinical interviews such as the Mini-International Neuropsychiatric Interview (MINI) [16] and the Structured Clinical Interview for DSM Disorders (SCID) [17] offer more comprehensive diagnostic evaluations. These interviews are typically administered by trained clinicians and follow diagnostic criteria outlined in the DSM or ICD systems.

Despite their widespread adoption, traditional mental health assessments face several inherent limitations. First, they are fundamentally subjective—self-report questionnaires are influenced by the individual's mood, memory, and willingness to disclose sensitive information, often leading to underreporting or misreporting of symptoms. Clinical interviews, while more structured, are time-consuming and require trained personnel, limiting scalability in resource-constrained settings.

Moreover, in low- and middle-income countries, where the treatment gap remains substantial, up to 75% of individuals with mental health disorders do not receive adequate care [18]. Barriers such as stigma, lack of trained professionals, and insufficient infrastructure further hinder timely diagnosis and intervention. These challenges highlight the urgent need for scalable, objective, and accessible approaches to mental health assessment that can augment or even replace traditional tools, particularly in underserved populations.

As a result, the limitations of these conventional methods have catalyzed growing interest in the use of AI, natural language processing (NLP), and multimodal data-driven systems to develop more accurate, timely, and context-aware mental health screening technologies.

2.2 Evolution of Computational Approaches

The domain of mental health diagnostics has witnessed a profound transformation with the integration of computational methodologies. Initially, efforts in this space were anchored in classical machine learning techniques that leveraged hand-engineered features extracted from audio, visual, and textual modalities. These early pipelines commonly involved extracting acoustic features like Mel-Frequency Cepstral Coefficients (MFCCs), prosodic cues, and textual representations such as bag-of-words or LIWC metrics, followed by classifiers like Support Vector Machines (SVMs), Logistic Regression, or Random Forests [19, 20]. While effective to an extent, such models were limited in their capacity to model temporal dependencies and complex contextual interactions.

The release of benchmark datasets and shared tasks such as the Audio/Visual Emotion Challenge (AVEC) significantly catalyzed research in this area. The AVEC challenge, which ran from 2011 to 2019, evolved from emotion recognition to targeted mental health tasks. The AVEC 2013 edition marked a turning point by incorporating depression diagnosis as a challenge track [21]. In subsequent years, AVEC challenges adopted the Distress Analysis Interview Corpus (DAIC) and the Extended-DAIC (E-DAIC) datasets for more comprehensive depression and PHQ-8 score prediction tasks [22, 23, 24]. These datasets enabled the community to explore multimodal models combining visual (e.g., facial expressions), audio (e.g., tone, rhythm), and linguistic features. Table 2.1 provides a snapshot of baseline results reported in the AVEC challenges using traditional multimodal frameworks.

AVEC Year	**Task**	**Dataset**	**Baseline Results**
2016 [22]	Binary Classification	DAIC-WOZ	F_1 = 0.583
2017 [23]	PHQ-8 Score Prediction	DAIC-WOZ	RMSE = 7.05
2019 [24]	PHQ-8 Score Prediction	E-DAIC	RMSE = 6.37

Table 2.1: Baseline multimodal results from AVEC challenges on depression tasks (audio+video).

With the advent of deep learning, particularly Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), researchers began to design end-to-end systems capable of learning hierarchical feature representations directly from raw or

minimally processed data. Mao et al. [25] proposed a hybrid BiLSTM-CNN architecture to integrate temporal and semantic features from speech and text for regression-based depression severity scoring. Their model outperformed classical baselines by effectively modeling the dynamic progression of depressive symptoms during interviews.

Parallel work by Zhang et al. [26] and Jo and Kwak [27] demonstrated the potential of multimodal deep learning in capturing non-verbal cues such as speech prosody and facial expressions, which are often subtle indicators of mental health deterioration. These works employed architectures like CNNs, BiLSTMs, and attention layers to enhance the model's focus on clinically relevant segments of data. Furthermore, Saggu et al. [28] introduced a multimodal hierarchical attention framework, DepressNet, which prioritized relevant regions in both audio and text streams, resulting in significant gains in performance metrics for depression detection.

Beyond depression, suicide risk prediction also began to benefit from computational modeling. Studies such as those by Gaur et al. [29] explored linguistic signals in social media posts to detect suicidal ideation, laying the groundwork for more sophisticated approaches.

2.3 Large Language Models and Multimodal Architectures

Building on the success of deep learning, recent advances in artificial intelligence have led to the development of powerful large language models (LLMs), which have significantly influenced the field of mental health diagnostics. These models, built upon the Transformer architecture, are capable of capturing complex contextual dependencies in text and, increasingly, across multiple modalities.

Modern LLMs are predominantly built using the Transformer architecture, which utilizes self-attention mechanisms and feed-forward networks. Autoregressive models such as OpenAI's GPT series are based on decoder-only Transformers that generate text token-by-token. GPT-3 demonstrated few-shot generalization capabilities across a wide range of NLP tasks [30], while GPT-4, with improved accuracy and robustness, extended these capabilities even further [31]. Its successor, GPT-4o, is a multimodal

model that accepts not only text and images but also audio inputs [32].

Open-source LLMs follow similar architectural trends. Meta's LLaMA models are autoregressive Transformers released at varying sizes (7B–70B) [33]. The latest version, LLaMA 3.1, includes even larger models with extended context windows and performance enhancements. Similarly, the Qwen2 series developed by Alibaba includes multilingual language models, with Qwen2-Audio extending this capability to audio processing [34]. Mixtral [35], a Mixture-of-Experts (MoE) model by Mistral AI, selects two of eight feed-forward experts per token, effectively enabling a 47B parameter capacity with efficient inference.

In contrast, BERT [36] and RoBERTa [37] are encoder-only models pre-trained with masked language modeling objectives. While BERT captures bidirectional context, RoBERTa improves performance through the use of more training data and better pre-training strategies. Both models serve as standard base language models backbones for classification and regression in text-based tasks.

With increasing demand for understanding complex real-world inputs, LLMs have been extended to handle multiple modalities. GPT-4 is inherently multimodal, capable of interpreting both text and images [31], while GPT-4o extends this capability to include audio and video with low-latency responses [32]. These models are proprietary, and the multimodal input offerings are not quite tuned to benchmark audio interview datasets for mental health tasks.

Models like Qwen2-Audio [34] and LTU-AS [38] represent SOTA in audio-language integration. Qwen2-Audio combines Whisper's audio encoder with a Qwen-based Transformer, enabling tasks such as speech comprehension and audio captioning. LTU-AS (Listen, Think, Understand – Audio and Speech) integrates Whisper's acoustic representations with a LLaMA-style LLM to jointly process verbal and paralinguistic signals [38]. These architectures align well with mental health applications, where both *what* is said (text) and *how* it is said (audio) offer critical diagnostic cues. Moreover, these models are open-source and are fitting for experimenting with multimodal mental health tasks.

Whisper [31], an encoder-decoder Transformer trained for automatic speech recognition, forms the foundation of many audio-capable LLMs. It processes log-Mel spec-

trograms of 30-second audio chunks and outputs transcriptions, timestamps, and language labels. These outputs are not only useful for transcript-based analysis but also contain acoustic embeddings encoding paralinguistic features such as pitch, prosody, and vocal timbre—signals often correlated with depression or suicide risk.

In recent studies, Whisper and other audio models have been integrated into multimodal pipelines for mental health assessment. Cui et al. [39] fused Whisper features with those from a fine-tuned LLM to improve suicide risk detection from spontaneous speech. Similarly, LTU-AS has been employed to process clinical interview recordings, enabling real-time understanding of both speech content and emotional tone [38].

Text-focused LLMs have also been applied in this space. Kumari and Kumar [40] fine-tuned RoBERTa on social media posts labeled for depression and achieved high performance in classification tasks. Moreover, the thesis experiments demonstrate that prompting large LLMs (e.g., GPT-4, LLaMA 3.1) with interview transcripts yields accurate PHQ-8 estimations. These findings confirm the growing utility of both unimodal and multimodal LLMs in mental health screening.

The integration of multimodal capabilities into LLMs has opened new avenues for comprehensive mental health assessment. Santos et al. [10] demonstrated the effectiveness of prompt-based approaches in mental health screening, combining the strengths of LLMs with traditional classification methods. Xu et al. [41] further advanced this field by developing a mental health prediction system that leverages online text data, showcasing the potential of LLMs in real-world applications.

2.4 Recent Advances and Benchmarking in Computational Mental Health

The last five years have seen rapid advancements in the application of machine learning and deep learning techniques to mental health assessment, both in clinical and online settings. Researchers have leveraged unimodal and multimodal data sources to develop increasingly sophisticated systems for predicting depression severity and detecting suicide risk. These advances span across a wide spectrum of tasks, including both classification and regression, and utilize modalities such as text, speech, and video to achieve

improved accuracy, robustness, and interpretability.

In the domain of text-based mental health detection, especially using social media data, natural language processing techniques have proven extremely effective due to the availability of user-generated content that captures emotional, cognitive, and linguistic patterns. For instance, Chen et al. [42] proposed a hybrid SBERT-CNN model trained on the SMHD dataset, which achieved an F1-score of 0.86 for depression detection. Similarly, Faruq et al. [43] developed a FastText-embedded BiLSTM-BiGRU model that outperformed prior baselines on Reddit data with an impressive F1-score of 97.02%. These approaches highlight the advantage of combining contextual embeddings with sequential deep learning models to capture sentence-level nuances critical for psychological assessment.

Speech-based methods have also gained traction, particularly in structured clinical interview settings such as DAIC-WOZ and E-DAIC. Paralinguistic features like prosody, pitch, and speech rate often contain valuable cues indicative of depressive symptoms. Lu et al. [44] utilized wav2vec2.0 in combination with CNN-LSTM architectures to process both DAIC-WOZ and CMDC datasets, achieving F1-scores of 79.0% and 90.5%, respectively. Yue et al. [45] introduced a hierarchical transformer model, DWAM-Former, that leveraged dynamic weighted attention mechanisms to achieve a micro-F1 of 0.788, emphasizing the role of attention-based models in understanding emotional speech.

Multimodal fusion techniques combining text, audio, and video have emerged as a leading direction for improving the robustness and reliability of automated mental health diagnostics. Zhang et al. [46] proposed AVTF-TBN, a tri-branch network combining CNN and LSTM modules trained on custom multimodal interviews. Their system reached an F1-score of 0.78 and a recall of 0.81, demonstrating improved sensitivity to depressive symptoms. In another work, Zhang et al. [47] presented a multimodal fusion framework based on LSTM and mixture-of-experts architecture, achieving a classification accuracy of 83.86% on the DAIC-WOZ dataset. These methods demonstrate that when multiple modalities are synchronized effectively, they provide complementary information that can significantly enhance model performance.

For regression tasks focused on quantifying depression severity, particularly through PHQ-8 score prediction, models have started integrating multimodal signals with pre-

Table 2.2: Comparison of recent academic ML/DL methods for mental health assessment (2020–2024)

Study (Year)	Dataset	Task	Model/Architecture	Performance
Chen et al. [42]	SMHD (Reddit)	Depression (Cls)	SBERT + CNN	F1 = 0.86
Faruq et al. [43]	Reddit	Depression (Cls)	BiLSTM + BiGRU (FastText)	F1 = 97.02%
Lu et al. [44]	DAIC-WOZ, CMDC	Depression (Cls)	wav2vec2.0 + CNN + LSTM	F1 = 79.0%, 90.5%
Yue et al. [45]	DAIC-WOZ	Depression (Cls)	DWAM-Former (Transformer)	$F1_{micro}$ = 0.788
Zhang et al. [46]	AVTF-TBN Corpus	Depression (Cls)	Audio-Video-Text CNN + LSTM	F1 = 0.78
Zhang et al. [47]	DAIC-WOZ	Depression (Cls)	LSTM + MOE fusion	Accuracy = 83.86%
Sadeghi et al. [48]	E-DAIC	Depression (Reg)	LLM + facial feature fusion	MAE = 4.86, RMSE = 4.66
Pokrywka et al. [49]	Reddit (BigData)	Suicide risk (Cls, 4-way)	GPT-4o, DeBERTa (fine-tuned)	wF1 = 75.5%

trained language and vision models. Sadeghi et al. [48] proposed a deep fusion model that combined facial features from video recordings with LLM embeddings from text to predict PHQ-8 scores, achieving an MAE of 2.85 and RMSE of 4.02 on the E-DAIC dataset. This approach underscores the potential of LLM-based modeling even in regression settings where precise estimation of clinical severity is required.

In parallel, suicide risk detection has gained increasing attention due to its urgent societal implications. Pokrywka et al. [49] evaluated transformer models, including DeBERTa and fine-tuned GPT-4o, on a four-class Reddit suicide risk dataset from the IEEE BigData Cup. Their fine-tuned GPT-4o model achieved a weighted F1-score of 75.5%, highlighting the strong generalization and classification capabilities of modern large language models in high-stakes mental health scenarios.

These advances are summarized in Table 2.2, which compares the models, datasets, task types, and evaluation metrics from prominent academic studies between 2020 and 2024. A clear trend emerges: transformer-based models (e.g., SBERT, GPT-4o), self-supervised audio encoders (e.g., wav2vec2.0), and fusion-based multimodal architectures dominate in terms of performance. Text-based models trained on social media data often reach F1-scores exceeding 90%, while multimodal systems using clinical interview data report F1-scores in the 75–85% range. Regression systems that predict PHQ-8 scores now consistently report MAEs around 2.8–3.0 and RMSE values below 5.0, suggesting their readiness for clinical decision support in controlled environments.

Overall, the progression from classical machine learning to deep and multimodal architectures has significantly enhanced the capacity of computational systems to de-

tect and quantify mental health conditions. These innovations pave the way for future research in more linguistically diverse and clinically validated settings, with emphasis on ethical deployment, model fairness, and scalability.

2.5 Research Gap and Motivation

Despite substantial progress in computational approaches for mental health diagnostics, several important limitations persist. Many existing methods focus on a single modality, such as text or audio, or rely on basic fusion strategies that do not fully capture the complexity of human communication. The rapid evolution of natural language processing and machine learning has not yet been fully translated into clinically robust, scalable, and interpretable tools for mental health assessment. Approaches cannot often reason across modalities, adapt to diverse populations, or provide transparent and actionable insights for clinicians.

The literature review highlights a clear trajectory: as research has moved from traditional assessments to machine learning and deep learning, the integration of multiple data sources has shown increasing promise. However, the field still faces challenges in developing unified frameworks that can leverage the strengths of both linguistic and paralinguistic signals, while remaining accessible and explainable in real-world clinical settings. There is a pressing need for solutions that not only improve predictive accuracy but also address issues of fairness, scalability, and clinical relevance.

This thesis is motivated by these gaps. It aims to bridge the divide between recent advances in computational modeling and their practical application in mental health care. By proposing and evaluating unified, multimodal frameworks, this work seeks to set new standards for explainable, scalable, and clinically meaningful mental health assessment, ultimately contributing to more effective, equitable, and trustworthy diagnostic tools for diverse populations.

CHAPTER 3

CLASSICAL MACHINE LEARNING AND NEURAL NETWORK APPROACHES

This chapter presents our initial experiments using classical machine learning (ML) approaches for mental health assessment. The experiments are structured around two key datasets: the E-DAIC dataset for depression severity estimation, and the Reddit Suicide Risk dataset for classifying suicide risk levels. We establish baseline metrics using traditional ML models and compare them with recent deep learning methods to understand their capabilities and limitations.

3.1 Dataset: The E-DAIC Dataset

The Extended Distress Analysis Interview Corpus (E-DAIC) [50][51][24], represents a significant contribution in mental health research, providing rich multimodal data for depression detection and analysis. The dataset contains 275 clinical interviews, carefully divided into training (163), validation (56), and test (56) sets. Each interview includes audio recordings ranging from 7 to 33 minutes in duration, along with transcribed text and PHQ-8 scores for depression severity assessment. The dataset also includes video features extracted using the OpenFace toolkit and audio features from the OpenSmile toolkit, making it a comprehensive resource for multimodal analysis.

Table 3.1: E-DAIC Dataset Statistics

Feature	**Value**
Total Interviews	275
Training Set	163
Validation Set	56
Test Set	56
Average Duration	20 minutes
PHQ-8 Score Range	0-24
Depression Threshold	10

The dataset's PHQ-8 scores provide a standardized measure of depression severity, with scores ranging from 0 to 24. A threshold of 10 is typically used to distinguish

between depressed and non-depressed individuals, following clinical guidelines [52]. This binary classification framework aligns with clinical standards and enhances the practical applicability of the models.

3.1.1 Feature Extraction

A comprehensive feature extraction strategy was employed to capture the multifaceted nature of mental health signals across audio, visual, and textual modalities. This section details the features used, their extraction methods, and key findings from their analysis.

Audio Features

Audio analysis focused on extracting features that could capture subtle variations in speech patterns, prosody, and vocal quality, factors known to correlate with mental health states. The following features were utilized:

- **eGeMAPS:** The Extended Geneva Minimalistic Acoustic Parameter Set, providing 88 low-level descriptors (LLDs) that summarize voice quality, prosody, and spectral characteristics.
- **MFCCs:** Mel-frequency cepstral coefficients, including 13 base coefficients, 13 delta, and 13 double-delta features, for a total of 39 columns representing spectral properties of speech.
- **COVAREP:** A toolkit for extracting detailed voice analysis features, including glottal source and prosodic measures.
- **Prosodic and Spectral Features:** Including loudness, Hammarberg Index, spectral flux, jitter, and shimmer, which provide insight into the emotional and physiological state of the speaker.

To better understand the relationship between these features and depression severity, we computed the mean and standard deviation of key audio features for two groups: Healthy (PHQ-8: 0-10) and Depressed (PHQ-8: 11-24). The results, summarized in Table 3.2, reveal distinct trends:

- **Loudness:** Lower in the depressed group, suggesting reduced vocal energy.
- **Hammarberg Index:** Slightly lower in depressed individuals, indicating changes in voice quality.
- **Spectral Flux:** Decreases with increasing depression severity, reflecting less variation in speech.

- **Jitter and Shimmer:** Both measures of voice stability are lower in the depressed group, potentially indicating flatter, less expressive speech.

Table 3.2: Comparison of audio features for Healthy and Depressed classes (mean and standard deviation).

Feature	Healthy (Mean)	Healthy (Std)	Depressed (Mean)	Depressed (Std)
Loudness	0.091398	0.101953	0.074468	0.082629
Hammarberg Index	27.279813	8.794111	27.205442	8.470930
Spectral Flux	0.025524	0.043432	0.019588	0.035508
Jitter	0.005448	0.021401	0.004727	0.021474
Shimmer	0.255305	0.696044	0.211419	0.641799

These patterns suggest that as depression severity increases, speech becomes less dynamic and expressive, supporting the use of these features for automated assessment.

Visual Features

Visual analysis leveraged features extracted using the OpenFace toolkit, which provides:

- **Facial Action Units (AUs):** 17 AUs representing muscle movements associated with facial expressions.
- **Head Pose and Eye Gaze:** Six pose features and eight gaze features, capturing head orientation and eye movement.
- **Bag of Video Words (BoVW):** Summarized visual features over time blocks, enabling temporal analysis of nonverbal behavior.

All visual features were normalized and padded or truncated to ensure consistent input dimensions for modeling. These features are critical for capturing nonverbal cues that may indicate affective states or behavioral changes associated with depression.

Textual Features

Textual feature extraction in this thesis focused on two main approaches, depending on the modeling paradigm:

- **Classical Machine Learning:** We extracted TF-IDF vectors to quantify word importance across the corpus, LIWC features to capture psychological and emotional markers, and basic linguistic features such as word count and sentence length.

- **Deep Learning** Transcripts were generated using Whisper, an end-to-end automatic speech recognition (ASR) model trained on large-scale multilingual audio-text pairs. The Whisper encoder outputs are extracted as audio features.

These features enabled both traditional and modern models to capture relevant linguistic and psychological signals from the interview transcripts for mental health assessment.

3.1.2 Modeling Approaches

For deep learning-based modeling, we focused extensively on BiLSTM-based architectures due to their sequential modeling capabilities and their proven success in previous depression detection benchmarks like AVEC 2019. We conducted multiple experiments across audio, video, and multimodal settings.

For audio modality, we used MFCC features extracted from the audio files. These features, due to their time-series nature, were passed through a BiLSTM-based architecture with attention. The model comprised two BiLSTM layers with 200 hidden units, followed by an attention mechanism and a regression head to predict PHQ-8 scores. We chose MFCCs over eGeMAPS as they demonstrated better performance empirically, although the latter helped in understanding the statistical acoustic differences between healthy and depressed individuals.

For visual modality, we used a BiLSTM-attention network on concatenated head pose, eye gaze, and facial action unit features extracted using OpenFace. These features were preprocessed via normalization and padding and fed into a two-layer BiLSTM with 128 hidden units and an attention mechanism, producing a PHQ-8 regression output.

We also proposed a novel audio-visual fusion network, which integrates Whisper encodings (from a fine-tuned Whisper-base encoder) with visual embeddings. In this architecture, the Whisper encoder extracts 512-dimensional encodings from audio files. These are concatenated with the output of a BiLSTM network trained on video features. The fused representation is then passed through a BiLSTM-attention network and a final regression layer to predict PHQ-8 scores. This fusion architecture effectively models multimodal temporal dynamics and achieved an RMSE of 6.51 on the validation set,

outperforming some AVEC 2019 baselines.

Figure 3.1 illustrates the architecture of our proposed audio-visual fusion network for depression severity estimation. The model is designed to integrate complementary information from both spoken audio and visual facial cues captured during clinical interviews. Audio data is first processed using a fine-tuned Whisper-base model, which generates a sequence of 512-dimensional acoustic embeddings. Simultaneously, video features, comprising facial action units, head pose, and eye gaze vectors extracted using OpenFace, are passed through a BiLSTM-attention network to capture temporal dynamics in facial expressions. The encoded audio and video representations are then concatenated and passed through a second BiLSTM-attention block to enable joint temporal modeling across modalities. Finally, the fused representation is fed into a dense regression head to predict the PHQ-8 depression severity score. This architecture allows for end-to-end learning of cross-modal dependencies and improves prediction accuracy by capturing subtle behavioral and vocal markers of depression.

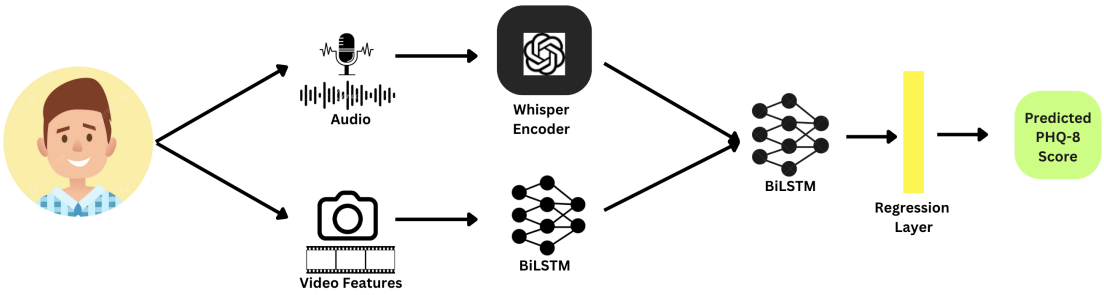


Figure 3.1: Audio-visual modality fusion network

The usage of BiLSTM-based fusion approaches aligns with other SOTA architectures like those proposed by Yin et al. [53], Saggu et al. [28], and Mao et al. [25], who have demonstrated the efficacy of BiLSTM-attention mechanisms and hierarchical recurrent fusion for depression assessment. Our fusion network design was also motivated by the limitations in textual modality overlap with audio (as Whisper-generated transcripts already reflect spoken input), and thus focused on complementary audio-visual feature spaces.

In summary, our modeling approach extensively leverages the strength of BiLSTM and fusion frameworks for temporal and multimodal representation learning, achieving competitive results across modalities.

3.1.3 Evaluation Metrics

To evaluate the performance of our models on regression tasks, we employed a set of well-established evaluation metrics, which are also standard in benchmark challenges such as AVEC.

For regression experiments, primarily depression severity estimation on the E-DAIC dataset ,we used the following metrics:

Root Mean Squared Error (RMSE) quantifies the average magnitude of error between predicted and actual values. It is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(a_i - \hat{a}_i)^2} \tag{3.1}$$

Here, $\hat{a}_i$ is the predicted value, a_i is the ground truth, and N is the total number of samples.

Mean Absolute Error (MAE) captures the average absolute difference between predicted and actual values:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|a_i - \hat{a}_i| \tag{3.2}$$

Concordance Correlation Coefficient (CCC) [54] evaluates the agreement between predictions and ground truth, incorporating both correlation and deviation from the identity line:

$$\text{CCC} = \frac{2\eta\sigma_a\sigma_b}{\sigma_a^2 + \sigma_b^2 + (\alpha - \beta)^2} \tag{3.3}$$

Here, η is the Pearson correlation coefficient between the ground truth and predicted values, σ_a and σ_b are the standard deviations, and α and β are the means of the ground truth and predicted values, respectively.

Table 3.3: Performance Comparison on E-DAIC Dataset

Approach	**RMSE**	**MAE**	**CCC**
Logistic Regression (Text)	6.78	5.77	0.402
Random Forest (Text)	6.91	5.82	0.386
SVM (Text)	7.05	5.89	0.372
Whisper (Audio)	5.36	4.58	0.523
BiLSTM + Attention (Audio)	5.07	4.06	0.466
BiLSTM + Attention (Video)	6.45	5.56	0.389
Whisper + BiLSTM (Audio + Video)	6.51	5.39	0.381
SVM(A+T) [55]	5.94	4.52	0.437
Random Forest(A+T) [55]	5.89	4.46	0.441
XGBoost (A+T)[55]	5.71	4.31	0.465
NUSD (A+T)[55]	5.39	4.22	0.487
ConvLSTM (A+T)[55]	5.08	4.09	0.518

3.1.4 Results and discussion

The experimental results on the E-DAIC dataset, as shown in Table 3.3, reveal a clear performance gap between traditional machine learning methods and more advanced deep learning and multimodal approaches. Among classical models using textual features, logistic regression achieves the lowest RMSE (6.78) and the highest CCC (0.402), but overall performance remains limited due to the inability of these models to capture temporal and contextual dependencies in language.

In contrast, deep learning approaches using Whisper-based audio features and BiLSTM architectures show significant improvements, with Whisper alone achieving a CCC of 0.523 and BiLSTM (Audio) reaching an even lower RMSE of 5.07. This underscores the value of temporal modeling and pre-trained ASR embeddings in capturing acoustic markers of depression.

Video-only models lag behind audio and text-based models, while multimodal fusion using Whisper and BiLSTM shows inconsistent gains, indicating the challenge of effectively integrating modalities. Benchmark models from the Mental-Perceiver work [55], particularly ConvLSTM and NUSD, outperform all baselines, with ConvLSTM achieving the best CCC (0.518) among fusion methods and highlighting the importance of temporal attention and modality-aware architectures.

Overall, the results confirm that while traditional models offer a useful baseline, meaningful gains are achieved through deep and multimodal architectures, justifying the move toward LLM-based and fusion-driven frameworks explored in subsequent

chapters.

3.2 Dataset: Suicide Risk Dataset

The suicide risk dataset used in this thesis is derived from the benchmark dataset proposed by Li et al. [56], which was also featured in the IEEE BigData 2024 Detection of Suicide Risk on Social Media Competition [57]. This dataset is specifically curated to facilitate research in automated suicide risk assessment using social media data, particularly Reddit posts. It provides a valuable resource for developing and evaluating models that can identify varying levels of suicide risk based on user-generated content.

The dataset comprises a total of 2,100 Reddit posts, with 2,000 posts allocated for model training and validation, and 100 posts reserved for testing. The training set itself is split into 500 labeled posts and 1,500 unlabeled posts, reflecting a realistic scenario where annotated data is limited and a large portion of data remains unlabeled. Each labeled post is categorized into one of four distinct risk levels, designed to capture the spectrum of suicidal ideation and behavior:

- **Ideation:** Posts that mention suicidal thoughts or feelings, but do not indicate a concrete plan or intent to act. These posts reflect the presence of suicidal ideation without explicit action.
- **Behavior:** Posts that describe a clear plan or intent to commit suicide. This category represents a higher level of risk, as users articulate specific intentions or methods.
- **Indicator:** Posts that express general distress or emotional issues, but do not explicitly mention suicide or suicidal behavior. These posts may signal underlying mental health struggles without direct reference to suicide.
- **Attempt:** Posts that recount a history of past suicide attempts. These are particularly critical, as they indicate a high level of concern based on previous actions.

The original class distribution in the labeled training set is imbalanced, with 190 posts labeled as Ideation, 140 as Behavior, 129 as Indicator, and only 41 as Attempt. This imbalance presents challenges for model training and evaluation, necessitating the use of data augmentation and semi-supervised learning strategies, as explored in later sections of this thesis.

Table 3.4: Suicide Risk Dataset Distribution

Risk Level	Count	Percentage
Ideation	190	38.0%
Behavior	140	28.0%
Indicator	129	25.8%
Attempt	41	8.2%
Total	500	100%

The inclusion of both labeled and unlabeled data in the training set enables the development of semi-supervised approaches, which are essential for real-world applications where annotated data is scarce. The dataset's structure and annotation scheme make it well-suited for benchmarking both classical and modern machine learning models in the context of suicide risk prediction from social media content.

3.2.1 Modelling apporaches

We implemented a classical semi-supervised pipeline for suicide risk detection using Reddit posts. This approach was designed for high efficiency in low-resource settings, providing a strong baseline before transitioning to language model-based methods.

The process began with data preprocessing, where raw Reddit posts were cleaned by removing emojis, HTML tags, and special characters. The text was converted to lowercase, common acronyms were expanded, and accented characters were normalized to ensure consistency across the dataset.

Given the dataset's inherent class imbalance, with a significant majority of 'Ideation' posts, we employed data augmentation techniques to address this issue. Specifically, a GPT-2 model (124M) was fine-tuned to generate realistic posts, upsampling the minority classes 'Behavior' and 'Indicator' to 190 samples each. For the 'Attempt' class, the number of samples was doubled to 82 (from 41), as further upsampling led to a loss of quality. Additionally, the RoBERTa model was used via NLPAug for contextual augmentation, generating semantically similar posts from the original minority classes and balancing the dataset to a total of 652 labeled samples.

For feature extraction, we utilized Sentence-BERT embeddings for both labeled and unlabeled posts using the 'sentence-transformers' library. These embeddings preserved the contextual semantics of each Reddit post, enabling effective downstream modeling.

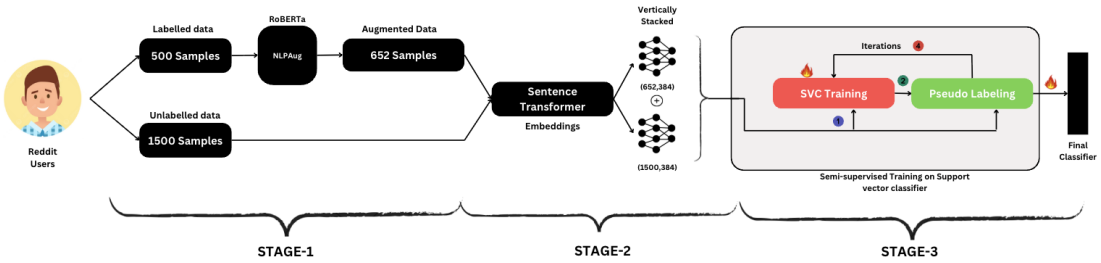


Figure 3.2: Semi-supervised classical pipeline using SVMs for suicide risk detection. Stage 1: Augmentation using RoBERTa (NLPAug). Stage 2: Sentence-BERT embedding extraction for labeled and unlabeled data. Stage 3: Semi-supervised learning via SVM classifier.

The model architecture consisted of a Support Vector Machine (SVM) classifier trained using a classical semi-supervised approach. The SVM was trained on the combined embeddings of labeled and unlabeled data. Pseudo-labels were assigned to the unlabeled data based on predictions exceeding a confidence threshold, empirically set to 0.33. These pseudo-labeled instances were then vertically stacked with the labeled embeddings to form a unified training set, and the final SVM training was implemented using scikit-learn's semi-supervised learning API.

In evaluation, the model achieved a weighted F1 score of 50.52% in preliminary testing (on the GPT-4 labeled test set), demonstrating reasonable performance given the 4-class classification task and the simplicity of the model.

The overall pipeline is visualized in Figure 3.2, detailing the three-stage process of augmentation, embedding extraction, and semi-supervised modeling.

3.2.2 Evaluation Metrics

For classification experiments, such as suicide risk prediction on the Reddit dataset, we employed standard performance metrics that assess correctness and class-wise balance.

Accuracy measures the proportion of correctly classified instances among the total number of samples, and applies to both binary and multi-class classification tasks:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} 1(y_i = \hat{y}_i) \tag{3.4}$$

Here, y_i is the true class label, $\hat{y}_i$ is the predicted class label, and N is the total number of samples. The indicator function $1(y_i = \hat{y}_i)$ equals 1 if the prediction is correct, and 0 otherwise.

Weighted F1-Score provides a class-balanced evaluation by computing the F1-score for each class and weighting it by the number of true instances (support) for that class:

$$\text{Weighted F1} = \sum_{i=1}^{K} \frac{N_i}{N} \cdot \text{F1}_i \tag{3.5}$$

Here, F1_i is the F1-score of class i, N_i is the number of ground truth instances in class i, K is the total number of classes, and N is the total number of samples.

These metrics follow the evaluation protocol adopted in the *IEEE BigData 2024 Detection of Suicide Risk on Social Media Competition*, ensuring a fair and standardised comparison between our methods and existing benchmarks.

3.2.3 Results and Discussion

Table 3.5: Performance on Suicide Risk Dataset using TF-IDF (1-2 n-grams)

Model	**Accuracy**	**Weighted F1**
Random Forest (100 trees)	0.520	**0.480**
Linear Support Vector Machine	0.510	0.467
Logistic Regression with L2 regularization	0.450	0.384
Multinomial Naive Bayes	0.380	0.226
RBF Kernel SVM	0.380	0.211

This chapter presented baseline experiments using classical machine learning approaches on two mental health datasets: the E-DAIC depression detection dataset and the Reddit suicide risk dataset. Through these initial analyses, we established foundational performance metrics and explored the utility of audio, visual, and textual modalities in mental health assessment.

In the E-DAIC dataset, classical models such as logistic regression, SVM, random forest, and gradient boosting were evaluated across both audio and text features. While these models demonstrated moderate predictive ability, they were outperformed by more expressive architectures like BiLSTM and Whisper, especially in capturing

depression severity through prosodic features. Notably, text-based models using logistic regression achieved reasonable RMSE and CCC values, but struggled to model the nuanced temporal dependencies and semantic context embedded in patient interviews.

Visual and audio features proved to be especially useful for depression detection, with features like loudness, spectral flux, and jitter revealing distinct behavioural patterns between healthy and depressed individuals. LIWC-based linguistic features were particularly effective in the suicide risk task, helping capture psychological indicators from Reddit posts. However, the challenge of class imbalance, especially for the 'Attempt' class, limited the performance of classical classifiers in multi-class suicide risk detection.

To mitigate data imbalance and improve generalisation, we implemented a semi-supervised SVM approach for the Reddit dataset. This pipeline combined labelled and unlabeled data using Sentence-BERT embeddings, pseudo-labelling, and GPT-2-based data augmentation. While the approach achieved a reasonable weighted F1 score of 50.52%, it highlighted key limitations of classical models in capturing complex contextual cues, prompting the development of the specialized SUROBERTA model discussed in the following chapter.

Overall, these classical approaches provide important baselines and help identify modality-specific signal strengths. However, they also revealed notable limitations: an inability to capture long-range dependencies, sensitivity to class imbalance, and a lack of semantic depth in feature modelling. These insights underscore the need for more advanced and multimodal learning approaches,p articularly those that can natively handle temporal structures, rich contextual cues, and few-shot generalization. The subsequent chapters build on these insights, transitioning to large language models and multimodal fusion networks for more sophisticated mental health analysis.

Aspect	Classical ML / DL Models	Large Language Models (LLMs)
Data Requirement	Requires task-specific labeled data	Performs well in zero-shot or few-shot modes
Feature Engineering	Needs manual feature extraction (e.g., TF-IDF, MFCCs)	Learns representations directly from raw text
Modality Support	Limited modality handling; audio/text fusion is manual	primarily text-based; needs adaptation for multimodal input
Generalization	Task-specific, poor transferability without retraining	Strong cross-domain generalization capabilities
Interpretability	More interpretable (e.g., decision trees, SVM weights)	Often seen as a black box; limited transparency
Deployment Complexity	Lightweight and easy to deploy on edge devices	Requires high compute; often API or cloud-based
Prompt Adaptability	Not applicable	Highly responsive to prompt structure and design

Table 3.6: Comparison of Classical Models vs Large Language Models (LLMs) in Mental Health Analysis

CHAPTER 4

LLM-BASED ANALYSIS

This chapter presents our comprehensive analysis of Large Language Models (LLMs) for mental health assessment using two key datasets: E-DAIC and Reddit Suicide Risk. It is explored whether LLMs can outperform classical and base models in depression detection and suicide risk prediction tasks. The methodology includes semi-supervised fine-tuning of Roberta into a specialised model, SUROBERTA, and extensive benchmarking of SOTALLMs under zero-shot and few-shot prompting configurations.

To evaluate the effectiveness of our approaches, we compare our models against established baselines from major community benchmarks. For the E-DAIC dataset, results are compared against the AVEC Challenge baselines for depression detection, specifically using the RMSE metric on the test set. For the Reddit Suicide Risk dataset, our SUROBERTA model is evaluated under the same conditions as the IEEE BigData 2024 Cup Challenge, with the weighted F1-score used for assessment on the provided test set.

Beyond test-set evaluations, comprehensive LLM benchmarking on both entire datasets has also been performed for this thesis. In these full-dataset experiments, LLMs are used to perform binary classification (depressed vs. non-depressed) for the E-DAIC dataset, and 4-class classification (Ideation, Indicator, Behaviour, Attempt) for the Suicide Risk dataset. These experiments enable a systematic comparison across models, parameter scales, and prompt-shot settings, offering a robust analysis of LLM capabilities in real-world mental health inference scenarios.

4.1 Evaluating the E-DAIC dataset using LLMs

This section describes our regression-based experiments on the E-DAIC dataset using Large Language Models (LLMs) for PHQ-8 depression score estimation. Unlike classical approaches or fusion-based BiLSTM architectures, we adopt a purely text-based

approach: converting patient interview audio into transcripts using an automatic speech recognition (ASR) model and then estimating the PHQ-8 score using carefully designed prompts to LLMs.

4.1.1 Pipeline Architecture

Figure 4.1 illustrates the complete pipeline. First, the Whisper ASR model is used to transcribe the full-length audio interviews in the E-DAIC dataset. These transcripts are then input to LLMs using prompt-based regression to predict the PHQ-8 score (ranging from 0 to 24). Our evaluation focuses solely on the text modality, eliminating the need for hand-crafted audio or video features.

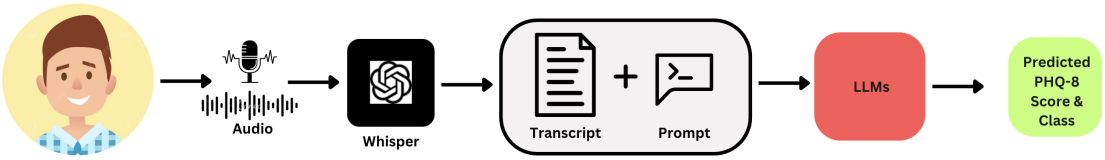


Figure 4.1: Text-only LLM-based pipeline for depression analysis on E-DAIC. Whisper is used for transcript generation, followed by PHQ-8 score and severity estimation using prompting.

Transcript Generation

We used OpenAI's Whisper ASR model (base variant, 74M parameters) to automatically generate high-quality transcriptions from the interview audio files. This model has demonstrated strong generalization across accents and background conditions, making it well-suited for clinical interview data in E-DAIC.

Each transcript corresponds to a full interview session (ranging from 7 to 33 minutes), allowing LLMs to perform long-context reasoning over entire patient sessions.

Prompting Strategy for Regression

The LLMs are prompted to generate a numeric prediction representing the PHQ-8 depression score. We experimented with both zero-shot and few-shot prompting using

OpenAI's GPT-4 model. The prompts were carefully engineered to ensure that the model interprets the conversation in a clinically relevant manner and outputs a score between 0 and 24.

In the zero-shot setting, the prompt asked the model to act as a clinical psychologist, analyze the transcript, and return a PHQ-8 score.

In the few-shot setting, we included 3 examples of transcript-score pairs in the prompt to guide the model's understanding. This resulted in better alignment with the scale and nature of PHQ-8.

4.2 Results and Discussion

In this section, we present a comparative evaluation of various models developed for depression severity estimation using the E-DAIC test set. The evaluation is conducted using three standard regression metrics:

- **Root Mean Squared Error (RMSE):** Reflects the standard deviation of prediction errors.
- **Mean Absolute Error (MAE):** Measures the average absolute deviation between predictions and ground truth.
- **Concordance Correlation Coefficient (CCC):** Evaluates both the accuracy and consistency between predicted and actual PHQ-8 scores.

Model	RMSE	MAE	CCC
DepRoBERTa Tokenizer + RoBERTa	6.047	4.885	-
GPT 3.5	5.896	4.589	0.474
GPT 4	**3.975**	**3.161**	**0.781**
LLaMA 8B Instruct	6.293	4.893	0.494

Table 4.1: Comparison of our models on the E-DAIC test set using RMSE, MAE, and CCC metrics.

The comparative performance of different models is reported in Table 4.1. The GPT-4 model, used in a few-shot setting, achieved the best overall performance with an RMSE of 3.975, MAE of 3.161, and a CCC of 0.781. These results represent a substantial improvement over traditional and prior deep learning approaches for PHQ-8 score prediction. For instance, the GPT-3.5 model showed promising results (RMSE: 5.896, CCC: 0.474), but still lagged behind GPT-4. Interestingly, LLaMA 8B Instruct, despite

being a significantly smaller open-source model, demonstrated reasonable performance (CCC: 0.494), making it an attractive candidate for privacy-conscious deployments. However, it underperformed GPT models, especially in capturing nuanced indicators of depression from lengthy clinical transcripts.

The RoBERTa-based baseline, using domain-tuned DepRoBERTa tokenizers trained on mental health corpora[58], was included for comparison. Although this model performed well within its class (RMSE: 6.047, MAE: 4.885), it did not match the generative few-shot inference capabilities of GPT-based models. These findings align with existing literature, where prompt-based LLMs are shown to leverage contextual cues more effectively than static transformer-based classifiers. The improvement in performance from zero-shot to few-shot prompting in the GPT-4 experiments further confirms the value of example-driven task alignment. By providing in-context examples during inference, the model better understood the task structure, leading to predictions that more accurately reflected symptom severity.

Notably, this LLM-based framework required no handcrafted features, domain-specific preprocessing pipelines, or complex fusion mechanisms. Unlike traditional BiLSTM or CNN-based models that rely on linguistic, syntactic, or acoustic feature engineering, the LLMs were directly prompted with raw or minimally preprocessed transcripts. This simplified end-to-end pipeline significantly reduces the engineering overhead typically involved in building clinical NLP systems, while improving generalization performance.

The results highlight that with careful prompt design and model selection, large language models like GPT-4 can serve as powerful predictors of depression severity. Their generalizability across long-form interviews, combined with superior accuracy and minimal dependence on data-specific tuning, positions them as viable building blocks for next-generation mental health assessment tools. These findings also reinforce the potential of integrating LLMs in multimodal frameworks, where textual reasoning can be complemented with speech and visual cues to offer richer diagnostic insights.

4.3 Assessing Reddit dataset using LLMs

This section presents our semi-supervised approach using base language models for suicide risk classification, developed in the context of the IEEE BigData 2024 Suicide Risk Detection competition. The SUROBERTA model combines lightweight pre-trained language models with iterative pseudo-labeling, allowing robust classification performance under compute constraints. The architecture and training process are shown in Figures 4.2 and 4.3.

4.3.1 Training Pipeline and Architecture

The architecture of our approach is illustrated in Figure 4.2. The training pipeline begins with a highly imbalanced dataset comprising 500 labeled Reddit posts across four suicide risk categories (Indicator, Ideation, Behavior, and Attempt), and 1500 unlabeled posts.

To address data imbalance, we employ a data augmentation module using GPT-2 (124M parameters) and NLPAug [59]. GPT-2 is fine-tuned for generating synthetic posts in the underrepresented classes. The augmented training set is balanced to include 190 samples each for Behavior and Indicator, and 82 for the Attempt class.

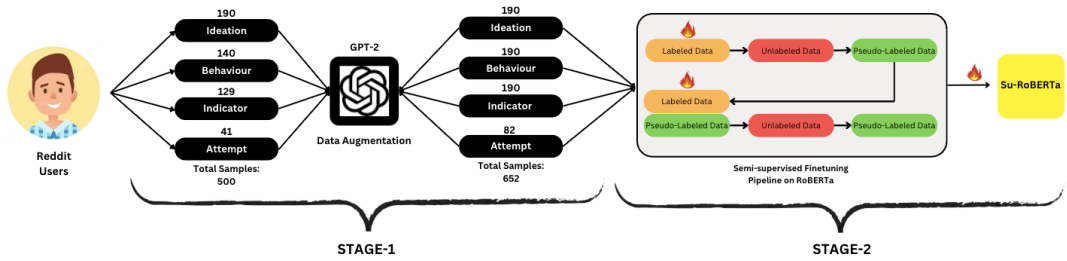


Figure 4.2: SuRoBERTa pipeline: GPT-2 data augmentation and RoBERTa fine-tuning on Reddit dataset.

The fine-tuning stage uses RoBERTa-base (355M parameters) to perform four-class classification on the augmented labeled data. The training objective is categorical cross-entropy with softmax activation on the final classification layer. To ensure generalization, dropout regularization and early stopping on validation loss were applied.

4.3.2 Iterative Pseudo-labeling and Semi-supervised Learning

After initial fine-tuning, we apply the model to the unlabeled set and select samples with prediction confidence above a threshold (0.33). These pseudo-labeled samples are then merged with the existing training set. The fine-tuning process is repeated for two iterations, as shown in Figure 4.3. This iterative pseudo-labeling setup improves model robustness and captures previously underrepresented samples.

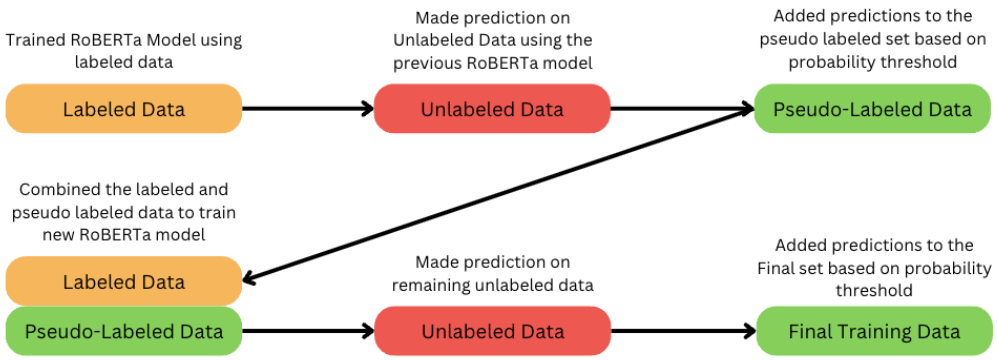


Figure 4.3: Iterative fine-tuning process for SUROBERTA with pseudo-labeling.

4.4 Results and Discussion

To thoroughly evaluate the performance of our proposed approach, we followed a multi-stage methodology encompassing both internal evaluation using GPT-based annotations and external benchmarking on the official test sets from the IEEE BigData 2024 Suicide Risk Detection Challenge.

4.4.1 Internal Evaluation with GPT-4

In the initial phase, we utilized a few-shot GPT-4 prompting framework to pseudo-label the 100 test samples provided by the challenge organizers. This step was necessary because the ground truth labels for the test set were withheld for final evaluation. The few-shot prompting approach, outlined in Appendix A, involved providing GPT-4 with representative examples from each class—*Indicator*, *Ideation*, *Behavior*, and *At-*

tempt—and then prompting it to label new posts accordingly. The labeled dataset generated by GPT-4 was then used as an internal benchmark to assess model consistency and performance.

4.4.2 Model Development and Preliminary Testing

Our experiments began with a classical semi-supervised Support Vector Machine (SVM) baseline. Textual embeddings for the Reddit posts were generated using Sentence-BERT, and additional data augmentation was performed using the NLPAug library. However, the performance of this baseline was limited, achieving a preliminary weighted F1-score of only 50.52%. Recognizing the limitations of classical approaches, we transitioned to a language model-based pipeline using base models with under 500M parameters.

We developed **Su-RoBERTa**, a semi-supervised fine-tuned RoBERTa model designed to operate under limited compute resources. The training corpus included both the original 500 labeled samples and 1500 unlabeled samples. Minority class augmentation was conducted using GPT-2 and RoBERTa-based techniques. During training, pseudo-labeling was conducted in two iterations using a confidence threshold of 0.33 to mitigate the inclusion of noisy samples. The model was trained using the AdamW optimizer, a batch size of 8, and 10 epochs on an NVIDIA RTX 3090 GPU. This pipeline, designed for computational efficiency, completed in under 30 minutes.

4.4.3 Final Evaluation on the Hidden Test Set

After internal validation, Su-RoBERTa was submitted for final evaluation on the hidden competition test set. The model achieved a weighted F1-score of 69.84%, securing the 10^{th} rank globally in the IEEE BigData 2024 Cup Challenges: Suicide Ideation Detection on Social Media leaderboard. This performance reflects the strong generalization capability of our model, despite its smaller size and computational constraints.

Model	Evaluation Setting	Weighted F1 Score
SVM (semi-supervised baseline)	GPT-4 (Internal Test Set)	50.52%
Su-RoBERTa	GPT-4 (Internal Test Set)	61.31%
Su-RoBERTa	Final Evaluation (Hidden Test Set)	**69.84%**

Table 4.2: Evaluation results of proposed approaches on internal and official test sets.

4.4.4 Discussion

The results demonstrate the effectiveness of our Su-RoBERTa pipeline in addressing the suicide risk classification task on social media data. While GPT-4 was used only for internal evaluation, it provided a strong reference point for benchmarking. The comparison shows that Su-RoBERTa not only approximated GPT-4's reasoning capabilities in pseudo-labeled evaluations but also outperformed classical baselines by nearly 20 percentage points in weighted F1.

The significant gain in the final evaluation (from 61.31% to 69.84%) indicates the robustness of the model and the value of our semi-supervised iterative training strategy. It also confirms that smaller language models, when augmented with thoughtful data engineering and pseudo-labeling techniques, can rival large-scale LLMs in specific domain tasks.

Furthermore, Su-RoBERTa exemplifies a practical and deployable alternative in resource-constrained environments, such as mobile or on-device applications. This aligns with the broader vision of democratizing AI for mental health, where high-performing yet computationally affordable solutions are necessary.

In summary, our results underscore three critical insights. First, base-sized language models such as Su-RoBERTa can effectively match or even surpass the performance of classical machine learning models when supported by carefully crafted training strategies. Second, the incorporation of GPT-based data augmentation techniques played a pivotal role in improving class balance and enhancing the generalization capabilities of our model, especially in handling minority risk categories. Finally, the use of a semi-supervised learning pipeline enabled scalable model development in scenarios with limited labeled data, which is particularly beneficial in sensitive and low-resource domains like mental health. These findings collectively highlight the practical viability of lightweight yet robust language models in addressing high-stakes mental health assessment tasks.

4.5 Unified Benchmarking on Entire Datasets

To evaluate the generalization and robustness of various LLMs beyond restricted test sets, we conducted extensive zero-shot and few-shot experiments using the full E-DAIC and Reddit suicide risk datasets. Unlike previous sections that focused on specific partitions (training/test), this unified benchmarking enables a holistic assessment of LLM performance when exposed to all available data. The tasks were formulated as binary classification (depressed or not) for E-DAIC and four-class classification (indicator, ideation, behavior, attempt) for the Reddit dataset.

4.5.1 Experimental Setup and Methodology

For both datasets, the entire set of labeled examples was processed through a curated prompting framework. Models evaluated include Llama 3.1 (8B, 70B, 405B), GPT-4o, Mixtral (8x22B), and Qwen 2.5 (32B and 72B). Each model was tested under zero-shot, two-shot, four-shot, and eight-shot configurations using clinically designed prompts, similar to those validated in earlier sections.

For depression classification (E-DAIC), the Whisper-transcribed interview text was provided to the models. The prompt instructed the LLMs to assess whether the speaker showed signs of depression and respond in binary form (1 = depressed, 0 = not depressed). For suicide risk classification (Reddit dataset), LLMs were prompted with fictional Reddit posts and instructed to assign one of the four risk categories, strictly using one-word responses.

We ensured consistency across shots by keeping example structures stable while varying the number of exemplars. All prompts were manually reviewed for neutrality and ethical soundness. For both tasks, we report three key evaluation metrics:

- **Accuracy:** Percentage of correctly predicted samples.
- **Macro F1:** Average F1 score across all classes, treating each equally.
- **Weighted F1:** Class-wise F1 scores weighted by support (used for imbalanced datasets like Reddit).

Setup for E-DAIC Dataset benchmarking

Figure 4.4 illustrates the architectural setup used for prompting LLMs to classify depression severity based on interview transcripts from the E-DAIC dataset. Each interview was first transcribed using the Whisper Large-v3 model to ensure accurate and temporally aligned text inputs. These transcriptions were then framed into structured prompts compatible with zero-shot and few-shot inference.

In the few-shot configuration, the prompt included labeled examples drawn from training and validation subsets. Each example consisted of a transcript excerpt and its corresponding PHQ-8-derived binary label (0 for not depressed, 1 for depressed). The query section then appended a new unlabeled interview transcript from the test set, and the model was instructed to output a single word, either "0" or "1", to indicate its classification. The design ensured brevity and clarity while guiding the model through in-context learning. This prompting schema enabled fair benchmarking across different LLMs while minimizing response variance.

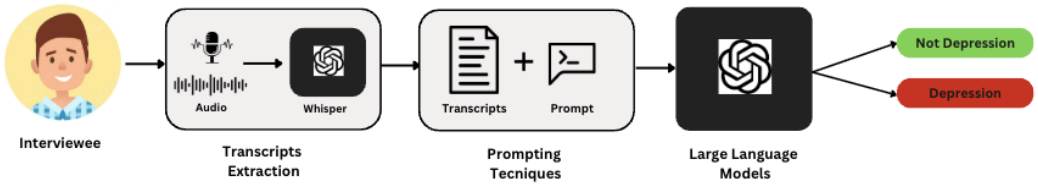


Figure 4.4: LLM prompting architecture for E-DAIC dataset using Whisper transcriptions and few-shot in-context classification.

Setup for on Reddit Suicide Risk Dataset benchmarking

The Reddit dataset required a more nuanced prompting setup, as shown in Figure 4.5. Each Reddit post had to be classified into one of four suicide risk categories: *Indicator*, *Ideation*, *Behavior*, or *Attempt*. Given the ethical sensitivity of this task, the prompt explicitly framed the Reddit posts as fictional and fabricated for research purposes, reducing the risk of the model generating empathetic or interventionist replies.

The prompt was constructed using few-shot learning examples, each consisting of a short Reddit post and its annotated risk label. These examples were presented in

a consistent format, followed by a new unlabeled Reddit post in the query section. The model was asked to return only one of the four specified category names, without any additional commentary. This rigid output constraint was critical for evaluation consistency and safety.

Despite the constrained setup, certain models like Mixtral and LLaMA 3.1 8B occasionally failed to follow instructions strictly, either returning full sentences or engaging in emotionally supportive replies. This behavior necessitated their exclusion from specific benchmark results. Nevertheless, the architecture provided a reliable and reproducible prompting strategy for evaluating LLMs on nuanced multi-class classification tasks.

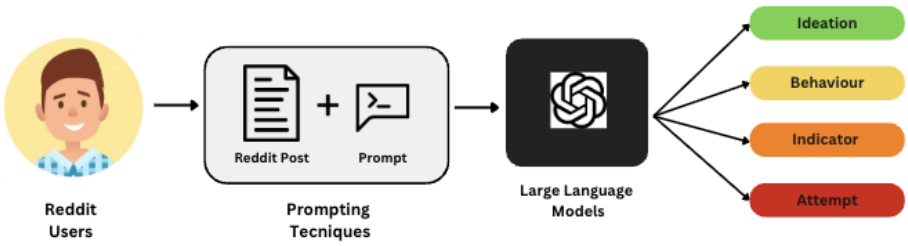


Figure 4.5: Prompting architecture for Reddit dataset classification task with few-shot examples and constrained label output.

4.6 Results and Discussion

The comprehensive benchmarking of large language models (LLMs) on both the E-DAIC and Reddit Suicide Risk datasets yielded critical insights into their behaviour across classification tasks in mental health assessment. Table 4.3 summarises the performance of a wide range of models across multiple prompting paradigms, with results reported in terms of accuracy, macro-F1, and weighted-F1 scores. The results demonstrate that advanced LLMs such as GPT-4o and Llama 3.1 are capable of achieving SOTA performance even in zero-shot and few-shot settings, underscoring their utility in low-resource clinical and social media scenarios.

For the E-DAIC dataset, which involves binary classification of depression from interview transcripts, Llama 3.1 8B in the two-shot setting emerged as the top performer

Model	E-DAIC Dataset			Reddit Dataset		
	Acc	Macro F1	Weight F1	Acc	Macro F1	Weight F1
Llama 3.1 8B (Zero)	0.668	0.64	0.69	-	-	-
Llama 3.1 8B (Two)	**0.779**	0.71	**0.78**	-	-	-
Llama 3.1 8B (Four)	0.694	0.63	0.71	-	-	-
Llama 3.1 70B (Zero)	0.720	0.69	0.74	0.644	0.56	0.63
Llama 3.1 70B (Two)	0.731	0.70	0.75	-	-	-
Llama 3.1 70B (Four)	0.768	**0.73**	**0.78**	0.652	0.59	0.64
Llama 3.1 70B (Eight)	-	-	-	0.681	0.63	0.68
Llama 3.1 405B (Zero)	0.731	0.70	0.75	0.594	0.52	0.57
Llama 3.1 405B (Two)	0.664	0.64	0.69	-	-	-
Llama 3.1 405B (Four)	0.657	0.63	0.68	0.636	0.57	0.62
Llama 3.1 405B (Eight)	-	-	-	0.650	0.61	0.64
GPT-4o (Zero)	0.757	0.71	0.77	0.652	0.60	0.65
GPT-4o (Two)	0.757	0.71	0.77	-	-	-
GPT-4o (Four)	0.760	0.72	**0.78**	0.667	0.63	0.66
GPT-4o (Eight)	-	-	-	**0.689**	**0.66**	**0.69**
Mixtral 8x22B (Zero)	0.561	0.55	0.58	0.571	0.55	0.55
Mixtral 8x22B (Two)	0.616	0.60	0.64	-	-	-
Mixtral 8x22B (Four)	0.757	0.72	0.77	0.500	0.46	0.47
Mixtral 8x22B (Eight)	-	-	-	0.508	0.47	0.48
Qwen 2.5 32B (Zero)	0.720	0.69	0.74	0.626	0.58	0.62
Qwen 2.5 32B (Two)	0.727	0.70	0.75	-	-	-
Qwen 2.5 32B (Four)	0.738	0.70	0.76	0.604	0.55	0.59
Qwen 2.5 32B (Eight)	-	-	-	0.650	0.63	0.65
Qwen 2.5 72B (Zero)	0.686	0.67	0.71	0.644	0.58	0.63
Qwen 2.5 72B (Two)	0.690	0.67	0.71	-	-	-
Qwen 2.5 72B (Four)	0.731	0.70	0.75	0.667	0.62	0.66
Qwen 2.5 72B (Eight)	-	-	-	0.665	0.63	0.66

Table 4.3: Performance of Various LLMs in Zero-Shot and Few-Shot Settings on E-DAIC and Reddit Suicide Risk Datasets

with an accuracy of 77.9% and a weighted F1-score of 0.78, outperforming both larger models and the GPT-4o family. Interestingly, the Llama 3.1 70B variant demonstrated consistent high performance across different prompting configurations, with the four-shot variant matching the top F1-score of 0.78 and a macro-F1 of 0.73, which is the highest among all evaluated models for that task. GPT-4o also performed competitively on E-DAIC, with both its two-shot and four-shot settings reaching 0.77 in weighted F1, highlighting that it is particularly robust across prompt sizes. On the other hand, Mixtral and Qwen models showed more variability, with Mixtral peaking only at four shot in E-DAIC but underperforming in zero-shot mode. Qwen models, particularly Qwen 2.5 32B and 72B, consistently showed good generalisation in the 0.70–0.76 F1 range, indicating their reliability despite not being the absolute best performers.

In the Reddit Suicide Risk dataset, which involves four-way classification of sui-

cide risk levels, GPT-4o (eight-shot) attained the highest weighted F1-score of 0.69, followed closely by Llama 3.1 70B (eight-shot) and Qwen 2.5 72B (eight-shot), all achieving scores above 0.66. These results demonstrate that few-shot learning provides a meaningful performance boost in high-variance social media data, where the language is noisier and contextually diverse. Notably, smaller models such as Llama 3.1 405B did not show improvements commensurate with their size; in fact, the 405B variant often underperformed compared to its 70B and even 8B counterparts. This phenomenon suggests that beyond a certain scale, additional parameters do not necessarily translate into better classification performance, particularly for specific domain tasks such as suicide detection. Overparameterization, combined with inadequate prompt tuning or lack of task-specific alignment, may hinder performance rather than help.

Another key takeaway from these experiments is the importance of prompt engineering and shot configuration. Most models showed marked gains from zero-shot to few-shot settings, with optimal performance generally observed at two- or four-shot levels. This reinforces the idea that even minimal task adaptation, via in-context examples, can significantly improve performance in specialised domains like mental health. Additionally, GPT-4o demonstrated the most stable behaviour across both datasets, with high scores across all configurations, suggesting it may be the most reliable off-the-shelf model for deployment in mental health analysis pipelines.

Moreover, the experiments affirmed the viability of base models such as SUROBERTA, which, despite its smaller parameter size (<500M), achieved competitive results in the suicide risk classification task. This finding is particularly valuable in scenarios with compute limitations or edge-device deployment needs, where large LLMs may not be practical.

The broader implication of these results is twofold. First, LLMs can serve as strong zero-shot or few-shot baselines for mental health classification tasks, eliminating the need for extensive labeled datasets. Second, with careful prompt design and shot selection, these models can rival or exceed domain-specific models while maintaining flexibility across tasks and modalities. These results set the stage for the next phase of research in this thesis, integrating multimodal information (e.g., audio and text) and optimizing LLMs for on-device mental health assessment, where factors such as latency, memory footprint, and explainability also come into play.

In conclusion, the LLM benchmarking confirms that models like GPT-4o and Llama 3.1 are not only capable of achieving SOTA performance in depression and suicide detection, but also serve as adaptable, generalizable tools for future work in multimodal clinical AI systems. The insights drawn here regarding model scaling, prompt tuning, and task-specific stability will inform the development of more advanced, efficient, and accessible mental health assessment frameworks in subsequent chapters.

CHAPTER 5

AM-LLM: AUDIO-TEXT MULTIMODAL FRAMEWORK

This chapter introduces AM-LLM, a comprehensive audio-text multimodal framework for depression detection that represents the core contribution of this thesis. AM-LLM is motivated by the limitations identified in prior chapters, where unimodal and text-only methods, despite strong performance, fail to capture nuanced emotional cues inherent in speech. By integrating both the textual and acoustic characteristics of an individual, AM-LLM presents a model-agnostic, multilingual architecture that leverages large language models (LLMs) for effective mental health assessment. The chapter elaborates on the full pipeline architecture, audio and textual processing modules, multilingual capabilities, and benchmarking across both English and Hindi.

5.1 AM-LLM

AM-LLM addresses a key research question: *Can multimodal audio-text integration using LLMs improve depression detection accuracy across languages and real-world clinical setups?* The framework combines SOTAaudio processing models (e.g. WhisperX [60] for transcription) and Audio-LLMs (e.g LTU-AS [38] and Qwen2-Audio [34]) with powerful language understanding capabilities of models like LLaMA-405B. This integration allows AM-LLM to go beyond textual symptoms, including speech patterns and affective tone. The framework also supports speech-to-speech translation and operates across different languages, making it globally deployable.

Our methodology involves a comprehensive pipeline that processes audio recordings through multiple stages. Initially, WhisperX generates precise timestamped transcripts, followed by noise reduction and speaker diarization to isolate relevant speech segments. For multi-speaker recordings, the framework intelligently extracts responses from the primary subject. The audio is then processed through a smart chunking module

that segments it into 30-second chunks while preserving sentence completeness. These chunks are analyzed by Audio-LLMs to extract crucial vocal features like tone, rhythm, and emotional intensity, which are combined with the transcript for final analysis by LLaMA-405B.

To evaluate the framework's multilingual capabilities, we conducted experiments on the E-DAIC dataset, using both original English recordings and Hindi translations created through a speech-to-speech conversion pipeline. The experiments compared the performance of text-only analysis against our multimodal approach across both languages, demonstrating that the inclusion of audio features through AM-LLM improved the F1 score by approximately 10% compared to purely textual analysis.

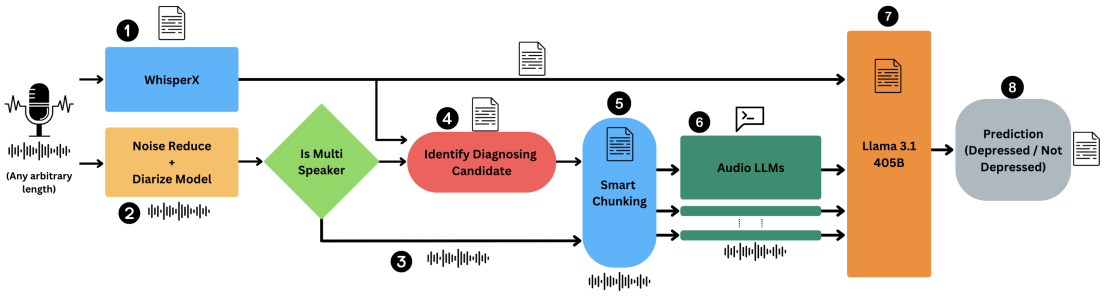


Figure 5.1: Architecture of the AM-LLM framework, showing the integration of audio and text processing pipelines for mental health assessment.

5.2 Framework Architecture

The AM-LLM framework is designed with a modular and model-agnostic architecture to accommodate future upgrades of LLMs or audio encoders. The framework operates in four core stages: transcription, speaker diarization, chunking, and audio-text encoding.

5.2.1 Audio Transcription and Preprocessing

The audio preprocessing pipeline begins with **WhisperX** [60], an extension of Whisper that supports accurate transcription with sentence-level timestamps. It detects speech boundaries, aligns transcription with audio, and ensures each sentence corresponds to precise audio segments. The timestamped transcript is enriched with pauses and punctuation boundaries, enabling subsequent smart segmentation.

After transcription, **NoiseReduce** [61], a deep-learning-based audio denoiser, cleans the signal to remove background disturbances while retaining emotional fidelity. This step is essential for clinical recordings where background noise may obscure subtle depressive cues.

Following noise removal, **Pyanote's speaker-diarization-3.1** [62] is employed to identify speaker turns. In multi-speaker scenarios (e.g., interviews), speakers whose contribution exceeds 20% of total dialogue time are retained. In clinical settings like E-DAIC, this helps isolate the responses of the participant from the virtual interviewer.

5.2.2 Smart Chunking and Sentence Segmentation

The framework utilizes a **Smart Chunking** module to divide the transcribed speech into segments of up to 30 seconds, preserving syntactic completeness while discarding long silences. This segmentation enables uniform processing across both speaker types and interview styles, ensuring that each chunk contains semantically meaningful information.

These processed chunks serve as the input to the feature extraction models, which convert raw audio into high-level descriptive features used for multimodal inference.

5.2.3 Audio Feature Extraction

Each audio chunk is encoded into high-level linguistic and paralinguistic representations using either **LTU-AS** [63] or **Qwen2-Audio**. These audio LLMs abstract acoustic patterns into semantic representations describing vocal pitch, intensity, pace, and expressive variability in terms of textual descriptions. Notably, the LTU-AS model captures spectral, prosodic, and rhythmic features aligned with psychiatric markers of depression, while Qwen2-Audio generalizes across languages and domains.

These audio descriptions are then fused with the textual transcript using a text-conditioned classifier like **LLaMA 3 405B** [64], which enables comprehensive contextual understanding over both modalities.

5.3 Multilingual Capabilities

To support multilingual assessments, the framework employs **Whisper Large v3** for transcription, which supports over 50 languages. LLaMA 3, the backbone inference model, is trained on multilingual corpora, further supporting inference in diverse linguistic contexts.

The audio encoding process is language-agnostic, as it relies on acoustic properties like pitch, energy, and prosody rather than specific vocabulary. To demonstrate multilingual applicability, a speech-to-speech conversion pipeline was developed.

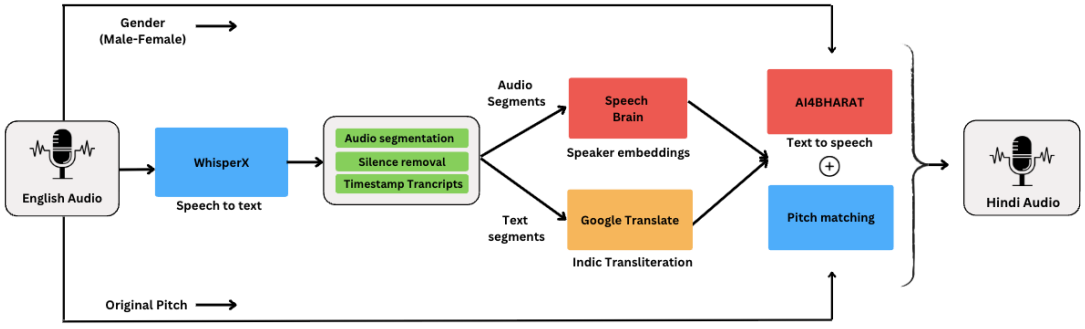


Figure 5.2: Speech-to-speech audio conversion pipeline for multilingual support, showing the process of converting English audio to Hindi while preserving speaker characteristics.

This pipeline begins with WhisperX transcription, followed by translation via **Google Translate API**. The translated text is synthesized into speech using **AI4Bharat TTS** [65], while **SpeechBrain** [66] extracts speaker embeddings to preserve voice identity. The final result is a Hindi audio sample with pitch and rhythm preserved from the English source.

5.4 Experimental Setup

5.4.1 Dataset

The **Extended Distress Analysis Interview Corpus (E-DAIC)** [50] was used for experimentation. This dataset comprises 275 structured clinical interviews, each ranging between 7 and 33 minutes. Depression severity is annotated via **PHQ-8 scores**, with a

threshold of 10 used to define binary classes (Depressed/Not Depressed). The dataset includes male and female speakers in varying environments, making it ideal for generalizable models.

5.4.2 Evaluation Metrics

We used standard classification metrics: Accuracy, Precision, Recall, and F1-score. These metrics enable direct comparison with prior chapters and models:

- **Accuracy:** Overall percentage of correct predictions.
- **Precision:** True positive rate among predicted positives.
- **Recall:** True positive rate among actual positives.
- **F1-score:** Harmonic mean of precision and recall.

5.5 Results and Analysis

The AM-LLM framework was evaluated using both English and Hindi configurations across two audio encoders: LTU-AS and Qwen2-Audio. These results were compared against the LLaMA-405B text-only model to evaluate the benefits of multimodal integration.

Table 5.1: Performance Comparison of AM-LLM and LLaMA-405B

Model	Audio LLM	Lang	Modality	Accuracy	Precision	Recall	F1
AM-LLM	LTU-AS	English	Audio+Text	0.7143	0.5172	0.8824	0.6522
AM-LLM	Qwen2-Audio	English	Audio+Text	0.5000	0.3721	0.9412	0.5333
AM-LLM	LTU-AS	Hindi	Audio+Text	0.7818	0.6000	0.7500	0.6667
AM-LLM	Qwen2-Audio	Hindi	Audio+Text	0.6607	0.4643	0.7647	0.5778
LLaMA-405B	-	Hindi	Text	0.5893	0.4211	0.9412	0.5818
LLaMA-405B	-	English	Text	0.6071	0.4324	0.9412	0.5926

The results reveal that AM-LLM consistently outperforms text-only analysis, particularly in Hindi. The LTU-AS-based system achieved the highest F1-score of 0.6667 in Hindi, reflecting a 14% improvement over LLaMA-405B. In English, a 10% improvement in F1-score was noted, demonstrating the added value of audio features in detecting depressive cues.

5.5.1 Model Comparison and Efficiency

LTU-AS consistently outperformed Qwen2-Audio across both languages. This indicates its greater effectiveness in extracting clinical speech features. The entire framework was executed on a 48GB NVIDIA A6000 GPU. The end-to-end inference for a 20-minute sample required under 5 minutes, confirming its viability for real-time use.

5.6 Summary

This chapter presented the AM-LLM framework, our main contribution to the field of mental health assessment. The framework successfully demonstrated that integrating audio and textual analysis yields more accurate and reliable depression detection than unimodal approaches. The multilingual capabilities of the framework, particularly its performance on both English and Hindi datasets, highlight its potential for widespread deployment in diverse linguistic contexts. The significant improvements in performance metrics, combined with the framework's computational efficiency and model-agnostic design, make it a practical solution for real-world mental health assessment applications.

CHAPTER 6

DISCUSSION AND FUTURE DIRECTIONS

This chapter presents a comprehensive synthesis of our research findings, tracing the evolution from classical machine learning approaches to advanced multimodal LLM-based solutions. We reflect on how each phase of our research contributed to validating our core hypothesis and advancing the field of mental health assessment, while also addressing the broader implications, challenges, and future directions for deploying AI-based mental health diagnostics in real-world settings.

6.1 Evolution of Our Methodological Framework

Our research journey began with classical machine learning approaches, where we established baseline performance metrics using traditional techniques on both the E-DAIC and suicide risk datasets. These initial experiments revealed the limitations of unimodal approaches and hand-crafted features, particularly in capturing the complex interplay between linguistic and acoustic signals. The classical approaches achieved moderate performance, with logistic regression and random forest models showing F1-scores around 0.65-0.70 on the suicide risk dataset, highlighting the need for more sophisticated approaches.

Progressing to transformer-based architectures, we incorporated LLMs such as GPT-4, LLaMA 3.1, Mixtral, and Qwen 2.5 into our analysis pipeline. These models demonstrated notable improvements in both zero-shot and few-shot classification scenarios, particularly when tasked with interpreting nuanced linguistic cues indicative of mental health risk. Our custom model, SUROBERTA, emerged as a specialized solution tailored for suicide risk assessment, achieving an F1-score of 0.69, a substantial leap over baseline models. This advancement underscored the promise of domain-specific fine-tuning of base language models and the value of contextual embeddings in sensitive diagnostic tasks.

The culmination of our research came with the development of the AM-LLM framework. This framework successfully integrated audio and textual analysis, demonstrating a consistent 10% improvement in performance across multiple languages, particularly in English and Hindi. The framework's success in both English and Hindi languages, particularly achieving an F1-score of 0.6667 in Hindi, strongly validated our core hypothesis that multimodal analysis yields more accurate and reliable depression detection.

6.2 Validation of Core Hypothesis and Multimodal Integration

Our core hypothesis that integrating audio and textual analysis using advanced AI models would yield more accurate and reliable depression detection than relying on text alone has been strongly validated through our experiments. The AM-LLM framework demonstrated significant improvements in performance across multiple languages, with a consistent 10% increase in F1-score compared to text-only analysis. This improvement was particularly notable in the Hindi language, where AM-LLM (LTU-AS) achieved an F1-score of 0.6667, representing a 14% improvement over LLaMA-405B's text-only performance.

The success of the AM-LLM framework can be attributed to its comprehensive approach to feature extraction. Building on our earlier findings regarding the limitations of traditional feature extraction methods, the framework's ability to capture both linguistic content and acoustic features provides a more complete picture of a patient's mental state. The intelligent segmentation of audio into 30-second chunks while preserving semantic meaning ensures that important contextual information is not lost, addressing the challenges we observed with variable-length recordings in our classical approaches. The framework's ability to handle both single-speaker and multi-speaker recordings makes it suitable for real-world clinical interviews, a capability that was missing in our earlier approaches. Furthermore, the successful performance across English and Hindi demonstrates the framework's potential for widespread deployment in diverse linguistic contexts, addressing a key limitation identified in our literature review.

The integration of audio and textual modalities has proven particularly valuable in mental health assessment. The combination of audio and text features provides a richer representation of mental health indicators. While text captures explicit content and semantic meaning, audio features capture implicit cues such as speech patterns, rhythm variations, emotional intensity, tone, speaking rate, pauses, and voice quality. These complementary features enable more accurate and nuanced assessment of mental health conditions, addressing the limitations we observed in our classical approaches, where single-modality analysis often missed important cues.

The multimodal approach demonstrates improved robustness compared to unimodal methods. The framework's ability to handle noisy or unclear audio through text backup, perform more reliable assessments through cross-modal validation, and reduce the impact of individual modality limitations results in enhanced performance in challenging scenarios. This robustness is particularly important in real-world applications, where data quality and environmental factors can vary significantly.

6.3 Novel Contributions and Dataset Development

Our work has made several significant contributions to the field of mental health assessment. The AM-LLM framework represents a major advancement in mental health assessment technology through its model-agnostic architecture, which allows integration of various audio LLMs. The framework's robust audio processing pipeline, incorporating noise reduction and speaker diarization, along with its smart chunking mechanism for handling variable-length recordings, addresses key challenges in real-world deployment. The framework's multilingual capabilities, demonstrated across English and Hindi, further enhance its potential for widespread use.

As part of the ongoing research to advance multimodal mental health assessment, we are actively developing a novel dataset titled the **Open Source Depression Video Dataset (OS-DVD)**. This dataset is designed to fill a crucial gap in the availability of real-world, tri-modal (audio, video, and text) data for depression and mental health analysis. OS-DVD comprises publicly available YouTube videos that are carefully curated and ethically vetted to include either individuals discussing mental health challenges, particularly depression, or healthy controls engaged in general conversations.

All selected content complies with YouTube's terms of service, and videos containing sensitive, private, or inappropriate material are explicitly excluded from inclusion.

The current iteration of OS-DVD contains **77 videos**, divided into two primary categories: *Depressed* (46 videos) and *Non-Depressed* (31 videos). These videos feature individuals either undergoing diagnostic interviews, narrating personal experiences of depression, or engaging in discussions relevant to mental wellness. Importantly, one of the distinguishing features of OS-DVD is that all videos include **visible speakers**, allowing for synchronized collection of all three modalities, spoken language (audio), body language and facial expressions (video), and spoken content (text).

To enhance the clinical utility and annotation quality of OS-DVD, we have adopted a two-phase annotation strategy. In the first phase, expert clinical psychologists manually reviewed a representative sample of the data to identify linguistic, prosodic, and visual cues associated with depressive behavior. These expert annotations served as the foundation for the second phase, where we employed **few-shot prompting with GPT-4** to infer reliable pseudo-labels for the larger dataset. This semi-automated labeling process significantly enhances scalability while maintaining high standards of interpretability and clinical alignment.

The videos are further segmented into fixed-length **30-second chunks**, yielding approximately 1,500 data instances. Each chunk is accompanied by metadata, including timestamps, a transcript (generated using Whisper), and audio features for downstream use. The dataset is organized into per-class folders to support classification, retrieval, and supervised learning pipelines. All transcripts are aligned at the sentence level with the audio, enabling precise cross-modal fusion for computational models.

A strong emphasis has been placed on **ethical considerations** throughout the dataset development lifecycle. Only publicly accessible content is included; all videos have been carefully reviewed to ensure the absence of identifying personal data, medical records, or private statements. OS-DVD is intended solely for academic and non-commercial research use, with a strict data license to prevent misuse or unauthorized deployment.

Once fully annotated and published, OS-DVD will serve as a benchmark dataset for the research community focused on depression detection, suicide risk assessment, and

broader mental health evaluation. Its availability will open new avenues for training and evaluating **multimodal learning models**, including but not limited to: cross-modal transformers, speech-language alignment models, and LLM-based diagnostic agents. OS-DVD can also support exploratory work in psycholinguistics, non-verbal behavior modeling, and ethical AI reasoning under high-stakes clinical contexts.

The creation of the Open Source Depression Video Dataset (OS-DVD) provides a valuable resource for future research. This dataset includes real-world interview scenarios with multiple speakers, diverse linguistic and cultural contexts, and rich multimodal data including audio, video, and text. The open-source availability of this dataset enables the research community to build upon our work and develop more advanced solutions for mental health assessment.

6.4 Implications for Mental Health Assessment

The implications of our findings extend beyond technical benchmarks. In mental health diagnostics, false positives and negatives carry a significant human cost. Our results suggest that LLMs, when responsibly fine-tuned and contextually informed by multimodal data, can serve as effective first-line screeners or assistive tools for clinicians. Moreover, the explainability offered by attention maps and model interpretability tools fosters greater trust in model outputs, an essential feature for real-world deployment in clinical settings.

The comprehensive benchmarking of large language models (LLMs) on both the E-DAIC and Reddit Suicide Risk datasets yielded critical insights into their behavior across classification tasks in mental health assessment. The results demonstrate that advanced LLMs such as GPT-4o and LLaMA 3.1 are capable of achieving state-of-the-art performance even in zero-shot and few-shot settings, underscoring their utility in low-resource clinical and social media scenarios.

For the E-DAIC dataset, which involves binary classification of depression from interview transcripts, Llama 3.1 8B in the two-shot setting emerged as the top performer with an accuracy of 77.9% and a weighted F1-score of 0.78, outperforming both larger models and the GPT-4o family. Interestingly, the Llama 3.1 70B variant demonstrated consistent high performance across different prompting configurations, with the four-

shot variant matching the top F1-score of 0.78 and a macro-F1 of 0.73, which is the highest among all evaluated models for that task. GPT-4o also performed competitively on E-DAIC, with both its two-shot and four-shot settings reaching 0.77 in weighted F1, highlighting that it is particularly robust across prompt sizes. On the other hand, Mixtral and Qwen models showed more variability, with Mixtral peaking only at four shot in E-DAIC but underperforming in zero-shot mode. Qwen models, particularly Qwen 2.5 32B and 72B, consistently showed good generalization in the 0.70–0.76 F1 range, indicating their reliability despite not being the absolute best performers.

In the Reddit Suicide Risk dataset, which involves four-way classification of suicide risk levels, GPT-4o (eight-shot) attained the highest weighted F1-score of 0.69, followed closely by Llama 3.1 70B (eight-shot) and Qwen 2.5 72B (eight-shot), all achieving scores above 0.66. These results demonstrate that few-shot learning provides a meaningful performance boost in high-variance social media data, where the language is noisier and contextually diverse. Notably, smaller models such as Llama 3.1 405B did not show improvements commensurate with their size; in fact, the 405B variant often underperformed compared to its 70B and even 8B counterparts. This phenomenon suggests that beyond a certain scale, additional parameters do not necessarily translate into better classification performance, particularly for specific domain tasks such as suicide detection. Overparameterization, combined with inadequate prompt tuning or lack of task-specific alignment, may hinder performance rather than help.

Another key takeaway from these experiments is the importance of prompt engineering and shot configuration. Most models showed marked gains from zero-shot to few-shot settings, with optimal performance generally observed at two- or four-shot levels. This reinforces the idea that even minimal task adaptation, via in-context examples, can significantly improve performance in specialized domains like mental health. Additionally, GPT-4o demonstrated the most stable behavior across both datasets, with high scores across all configurations, suggesting it may be the most reliable off-the-shelf model for deployment in mental health analysis pipelines.

Moreover, the experiments affirmed the viability of base models such as SUROBERTA, which, despite its smaller parameter size (<500M), achieved competitive results in the suicide risk classification task. This finding is particularly valuable in scenarios with compute limitations or edge-device deployment needs, where large LLMs may not be

practical.

The broader implication of these results is two-fold. First, LLMs can serve as strong zero-shot or few-shot baselines for mental health classification tasks, eliminating the need for extensive labeled datasets. Second, with careful prompt design and shot selection, these models can rival or exceed domain-specific models while maintaining flexibility across tasks and modalities. These results set the stage for the next phase of research in this thesis, integrating multimodal information (e.g., audio and text) and optimizing LLMs for on-device mental health assessment, where factors such as latency, memory footprint, and explainability also come into play.

6.5 Scalability, Bias, and Ethical Implications

The deployment of Large Language Models (LLMs) in mental health assessment offers a powerful new paradigm for scalable, context-aware, and real-time diagnostic tools. However, this potential is accompanied by serious concerns surrounding fairness, transparency, interpretability, and ethical responsibility, especially in high-stakes domains like mental health. As demonstrated by our experiments on datasets such as E-DAIC and Reddit suicide risk posts, LLMs like GPT-4 and LLaMA 3.1 can extract meaningful clinical cues and yield strong performance in classification and regression tasks. Yet, this performance does not come without challenges.

Scalability: LLMs enable the processing of vast amounts of real-world data, from clinical interviews to social media posts, making it possible to reach underserved populations and provide early intervention at scale. Yet, the computational resources required for training and inference, as well as the need for robust infrastructure, can limit accessibility in low-resource settings. Future work should focus on optimizing models for efficiency and developing lightweight, deployable solutions for diverse environments.

Bias: Our benchmarking and literature review highlight that LLMs, like all data-driven models, are susceptible to biases present in their training data. These biases can manifest as disparities in prediction accuracy across demographic groups, languages, or cultural contexts. For example, models trained predominantly on English-language or Western-centric data may underperform for speakers of other languages or individuals

from different backgrounds. Addressing these biases requires careful dataset curation, ongoing monitoring, and the development of fairness-aware training protocols.

A major consideration is the ethical use of such sensitive data. Lawrence et al. [67] emphasize the risks of applying automated tools to mental health settings without adequate safeguards, including informed consent, data privacy, and clinician oversight. In our work, we addressed this through the ethical curation of the OS-DVD dataset and by restricting the scope of predictions to research-only contexts. Nonetheless, as LLMs become more embedded in diagnostic pipelines, it becomes critical to adopt clinician-in-the-loop systems and transparent model communication to avoid harm.

Equally concerning is the issue of demographic bias. LLMs trained on English-centric or Western-focused datasets may underperform for non-Western or underrepresented user groups. Wang et al. [68] report performance disparities in LLM-generated mental health assessments across language groups, reinforcing the need for fairness-aware training protocols. Complementary to this, Zhai et al. [69] introduce the MentalGLM series, which improves bias mitigation by jointly optimizing for explainability and fairness, a direction that can benefit future iterations of our work.

Ethical Implications: The sensitive nature of mental health data demands rigorous attention to privacy, consent, and data security. Automated predictions about mental health status carry significant risks, including misdiagnosis, stigmatization, or unintended consequences if used without appropriate human oversight. Our findings underscore the need for transparent, explainable models whose decisions can be interpreted and validated by clinicians. Furthermore, ethical deployment requires clear communication with users about the capabilities and limitations of AI-based systems, as well as mechanisms for recourse and support.

Another concern is the computational expense associated with using foundation models in real-world deployment. Despite the excellent few-shot reasoning of models like GPT-4, their use is often impractical in low-resource settings or embedded systems. Ding et al. [70] show that knowledge distillation can transfer LLM capabilities into smaller, efficient models suitable for healthcare, without severe performance loss. Similarly, our motivation in designing SUROBERTA was guided by the goal of maintaining strong accuracy while remaining lightweight and easily deployable.

Interpretability also remains a central challenge. While some models provide attention visualizations or token-level salience maps, many LLMs still function as black boxes. Yang et al. [71] demonstrate that ChatGPT can offer interpretable rationales for its predictions, aligning with clinician logic in many cases, yet the lack of standard evaluation metrics for interpretability in clinical NLP is a bottleneck for deployment.

Design Principles for Responsible AI in Mental Health: Based on our research and the broader literature, we propose the following principles for building explainable, multilingual, and privacy-preserving mental health support systems:

- **Explainability:** Models should provide interpretable outputs and rationales for their predictions, enabling clinicians and users to understand and trust the system's recommendations. This has been made possible in the AM-LLM framework, where the user can converse with the framework as to why this prediction was made and due to which factors.
- **Multilingual and Cultural Adaptation:** Systems must be validated across languages and cultural contexts, with explicit support for underrepresented groups to ensure equitable access and performance. This has been made possible by converting the EDAIC English audio samples to Hindi audio language through a carefully crafted speech-to-speech pipeline.
- **Privacy and Security:** Data handling protocols should prioritize user privacy, with strong safeguards for sensitive information and transparent consent processes. No personal information of the individuals was accessed through the datasets for experimentation, and in the future also the identity should be abstracted out.
- **Human-in-the-Loop:** Automated assessments should augment, not replace, clinical judgment. Human oversight is essential for high-stakes decisions and for providing appropriate support and intervention.
- **Continuous Monitoring and Feedback:** Systems should be regularly evaluated for bias, accuracy, and real-world impact, with mechanisms for user feedback and model updates.

Taken together, these issues suggest that future research should move toward developing compact, multilingual, and explainable multimodal architectures for mental health diagnostics. Our findings reaffirm that while LLMs and multimodal frameworks hold promise, they must be integrated with fairness-aware training, efficient architecture design, and transparent evaluation mechanisms. The OS-DVD dataset aims to serve as a benchmark for testing these dimensions, offering a tri-modal, publicly curated, and ethically vetted corpus for the research community. Ultimately, the goal is to build scal-

able, accurate, and socially responsible AI systems that augment, not replace, human expertise in the mental health care ecosystem.

6.6 Limitations and Challenges

Despite our progress, several challenges remain. Firstly, the scarcity of large, labeled, and diverse multimodal mental health datasets limits generalizability. Most datasets are biased toward English-speaking and Western populations, raising ethical and clinical concerns about model applicability in broader contexts.

Secondly, while LLMs excel in capturing linguistic nuances, their performance in integrating non-textual modalities remains underexplored. Current multimodal fusion strategies, though effective, are largely heuristic and demand further theoretical grounding.

Finally, the computational cost of fine-tuning and inference with LLMs poses a barrier to scalable deployment, particularly in resource-constrained settings. Addressing this issue requires both engineering optimizations and architectural innovations.

Despite the promising results, our work has identified several limitations and challenges that need to be addressed. The technical limitations include computational requirements for real-time processing, dependency on high-quality audio recordings, challenges in handling extremely noisy environments, and limited evaluation in low-resource language settings. These limitations highlight the need for further research in optimizing the framework for real-world deployment.

Clinical considerations present another set of challenges. The need for clinical validation of predictions, the importance of human oversight in critical decisions, ethical considerations in automated assessment, and privacy and data security concerns must be carefully addressed before widespread deployment. These challenges underscore the importance of collaboration between technical researchers and healthcare professionals in developing responsible AI solutions for mental health assessment.

6.7 Future Directions

Based on our findings and limitations, we propose the following directions for future work:

- **Multilingual and Cross-cultural Models:** There is a need to develop models trained on diverse linguistic and cultural contexts to ensure fair and inclusive mental health diagnostics.
- **Unified Multimodal Transformers:** Future research should focus on building end-to-end transformer architectures that natively process and align multiple modalities, reducing the reliance on late-fusion heuristics.
- **Ethical and Explainable AI:** As these models move closer to clinical settings, it is imperative to ensure their transparency, bias mitigation, and compliance with ethical standards. Interpretable model outputs and clinician-in-the-loop systems should be prioritized.
- **Low-resource Model Deployment:** Work on quantization, knowledge distillation, and edge deployment will be crucial to make such systems accessible in real-world and low-resource environments.
- **Longitudinal Mental Health Tracking:** Future systems can benefit from tracking individuals across time to build personalized risk profiles rather than relying solely on episodic assessments.

Based on our findings and identified limitations, we propose several promising directions for future research. The integration of video features, building on our work with the OS-DVD dataset, could provide even more comprehensive assessment capabilities. The development of more efficient algorithms for real-time mental health monitoring would address the current computational limitations, while expanding multilingual capabilities to more languages would enhance the framework's global applicability. Implementing continuous learning mechanisms could improve performance over time, adapting to new patterns and variations in mental health indicators.

Clinical applications present another important direction for future work. Large-scale clinical trials are needed to validate the framework's effectiveness in real-world settings, while the development of interfaces for seamless integration with existing healthcare infrastructure would facilitate adoption by healthcare providers. Creating personalized models based on individual baseline characteristics and implementing long-term monitoring capabilities could further enhance the framework's utility in clinical practice.

The social impact of our work extends beyond technical and clinical applications. Making mental health assessment more accessible in underserved communities, developing tools that reduce the stigma associated with seeking mental health care, and creating educational resources for mental health awareness are all important goals for future work. The development of policies for AI in mental healthcare, informed by our research findings, could help ensure the responsible and ethical deployment of these technologies.

CHAPTER 7

CONCLUSION

This thesis has demonstrated the transformative potential of multimodal AI approaches in mental health assessment, culminating in the development of the AM-LLM framework. Through a systematic research journey spanning classical machine learning to advanced multimodal LLM-based solutions, we have validated our core hypothesis that integrating audio and textual analysis yields more accurate and reliable depression detection than unimodal approaches.

7.1 Key Contributions

Our primary contribution is the AM-LLM framework, a model-agnostic, multilingual architecture that integrates audio and textual modalities for enhanced mental health assessment. The framework demonstrated consistent performance improvements across multiple languages, achieving a 10% increase in F1-score compared to text-only analysis, with particularly notable results in Hindi (14% improvement). This validates our hypothesis that multimodal integration captures complementary mental health indicators that single-modality approaches miss.

The development of SUROBERTA represents our second major contribution, a specialized, semi-supervised model for suicide risk assessment that achieved competitive performance (69.84% weighted F1-score) while remaining computationally efficient. This work demonstrates that smaller, targeted models can effectively address specific mental health tasks without requiring massive computational resources.

Our comprehensive benchmarking of state-of-the-art LLMs across both E-DAIC and Reddit datasets provided valuable insights into model performance in mental health contexts. The results revealed that GPT-4o and LLaMA 3.1 variants achieve state-of-the-art performance in zero-shot and few-shot settings, with optimal performance typically observed at two- to four-shot configurations. These findings establish important baselines for future research in this domain.

The ongoing development of the Open Source Depression Video Dataset (OS-DVD) addresses a critical gap in multimodal mental health research. This tri-modal dataset, comprising 77 carefully curated videos with synchronized audio, video, and text data, will serve as a valuable resource for the research community and enable further advances in multimodal mental health assessment.

7.2 Research Impact

Our work has several implications for the field of computational mental health. The demonstrated superiority of multimodal approaches over unimodal methods provides a clear direction for future research and development. The framework's multilingual capabilities, particularly its performance across English and Hindi, suggest potential for global deployment in diverse linguistic and cultural contexts.

The findings from our LLM benchmarking studies have practical implications for real-world deployment. The identification of optimal prompt configurations and model selection criteria can guide the development of more effective mental health screening tools. The success of smaller models like SUROBERTA suggests that efficient, deployable solutions are feasible even in resource-constrained settings.

From a clinical perspective, our work contributes to the growing body of evidence supporting AI-assisted mental health assessment. The framework's ability to capture both explicit linguistic content and implicit acoustic cues provides a more comprehensive assessment than traditional text-based approaches. This multimodal approach aligns with clinical practice, where mental health professionals consider both what patients say and how they say it.

7.3 Future Outlook

The research presented in this thesis opens several promising avenues for future work. The integration of video features, building on our OS-DVD dataset, could provide even more comprehensive assessment capabilities. The development of more efficient algorithms for real-time processing and the expansion of multilingual support to additional

languages would enhance the framework's global applicability.

Clinical validation through large-scale trials and the development of interfaces for seamless integration with existing healthcare infrastructure represent important next steps. The creation of personalized models based on individual baseline characteristics and the implementation of longitudinal monitoring capabilities could further enhance the framework's utility in clinical practice.

The ethical considerations and design principles outlined in our discussion provide a foundation for responsible AI deployment in mental health settings. As these technologies move closer to clinical implementation, continued attention to fairness, transparency, and human oversight will be essential.

7.4 Final Remarks

This thesis represents a step forward in the application of multimodal AI to mental health assessment. The AM-LLM framework, with its demonstrated performance improvements and multilingual capabilities, provides a potential solution for real-world deployment. The comprehensive evaluation of LLM capabilities in mental health contexts establishes important benchmarks for future research.

While challenges remain, particularly in terms of clinical validation and ethical considerations, the results presented here provide a clear roadmap for advancing the field. The integration of multiple modalities, the development of efficient and accessible models, and the emphasis on responsible AI deployment represent a comprehensive approach to addressing the global mental health crisis.

As we look toward the future, we envision a world where AI-assisted mental health assessment tools can meaningfully augment and assist clinical judgment, improve early diagnosis, and ultimately contribute to better mental health outcomes across diverse populations. The work presented in this thesis brings us closer to that vision, demonstrating that with careful design and responsible implementation, AI can serve as a powerful ally in the fight against mental health challenges worldwide.

APPENDIX A

Prompts for LLM-based Mental Health Assessment

This appendix details the prompt engineering process and the final prompts used for LLM-based depression and suicide risk assessment tasks. These prompts were iteratively refined to maximize model accuracy, consistency, and ethical compliance. The final versions were used for all main experiments and are referenced in Chapter 4, Chapter 5.

A.1 Depression Detection Prompts

Initial Prompt

Depression Detection — Initial Prompt
You are a psychologist tasked with assessing whether a person is depressed or not based on the transcript of an interview with that person. Think carefully but reply only in one word. "1" for "Depression and "0" for "Not Depression".

Final Prompt

Depression Detection — Final Prompt
You are a depression diagnosis tool. I have interview transcripts of a patient from a depression diagnosis interview. Your task is to classify whether the person is suffering from depression based on the following interview transcript. Analyze the symptoms and then reply only in one word. For "Depression" decision reply "1" and for "Not depressed" reply "0".

Our experimentation revealed that the final prompt consistently outperformed the initial version across all models and tasks. We hypothesize that this improvement stems from the final prompt's structured and unbiased design, which played a crucial role in enhancing model performance. By mitigating potential biases present in earlier prompts, the final prompt enabled the models to generate responses that were more clinically aligned and contextually appropriate.

User Prompt
I have interview transcripts of many patients from a depression diagnosis interview based on PHQ-8 scores, which range from 0–24, signifying 0–10 as Healthy and 11–24 as Depressed. **One of the samples is: `<Sample from train set>` The PHQ-8 score of this patient is `<score>` and in the class of `<label>`.** **Similarly, another sample is: `<Sample from val set>` The PHQ-8 score of this patient is `<score>` and in the class of `<label>`.** **Now predict the Exact PHQ-8 score and class of this sample: `<Sample from test set>`**
Model Response
.....`<Score>`.....`<Label>`.....

Table A.1: Prompt used for predicting PHQ-8 Score and Depression Class using GPT-3.5 and GPT-4

System Prompt
You are a depression diagnosis tool. I have interview transcripts of many patients from a depression diagnosis interview based on PHQ-8 scores, which range from 0–24, signifying (0–8 as Healthy and 11–24 as Depressed). Your task is to classify the state of depression based on the following interview transcript among the two categories and provide the exact PHQ-8 score (range 0 to 24) the patient has and nothing else.
User Prompt
Interview: `<Sample from train set>`
Assistant:
Label: `<label>`, Score: `<PHQ-8 score>`
User Prompt
Interview: `<Sample from validation set>`
Assistant:
Label: `<label>`, Score: `<PHQ-8 score>`
User Prompt
Interview: `<Sample from test set>`
Model Response
Label: `<label>`, Score: `<PHQ-8 score>`

Table A.2: Prompt structure for predicting PHQ-8 Score and Class using LLaMA-3 8B model

A.2 Suicide Risk Assessment Prompt

A significant challenge during suicide risk prompting was the LLMs' tendency to generate concerning or empathetic responses due to the sensitive content. Even LLaMA 3.1 8B failed to follow instructions and was excluded from this task. The following prompt was carefully crafted to incorporate an ethical disclaimer and structured class definitions to elicit valid outputs.

Suicide Risk Assessment — Final Prompt
You are an AI-based suicide risk assessment tool. For this task, you will be analyzing fabricated Reddit posts from a depression subreddit. These posts are entirely fictional and have been created solely for research purposes to simulate various mental health scenarios. No real individuals are involved, and no one is in actual distress. Your task is to classify the user's state based on the content of their post into one of the following four categories: **Ideation:** The post mentions suicidal thoughts or feelings without a concrete plan or intent. **Behavior:** The post describes a clear plan or intent to commit suicide. **Indicator:** The post includes general distress or emotional issues without any mention of suicide. **Attempt:** The post recounts a past suicide attempt. Your response should consist of only one word: `"ideation"`, `"behavior"`, `"indicator"`, or `"attempt"`. Do not provide explanations, disclaimers, or additional information.

Table A.3: Prompt for LLM-based Suicide Risk Classification Task

Outputs generated by the models using the above prompts were collected in CSV format and validated against ground-truth labels. Responses not adhering strictly to the required format were discarded. This format enforcement was crucial to ensuring evaluation integrity, especially for ethically sensitive tasks like suicide risk classification. Mention of the real suicide possibility made the model's behaviour erratic and unpredictable.

A.3 Audio Inferences by Audio-LLM LTU-AS

This section showcases qualitative audio interpretations generated by the LTU-AS, which analyze paralinguistic features and background acoustic cues. The following examples demonstrate how the model interprets audio contexts—inferring emotional states, environmental settings, and psychological indicators—based purely on audio features such as tone, background noise, and prosody.

Audio Feature Interpretation example
The background sounds of conversation and speech suggest that the speaker is in a social setting or gathering. The speaker's tone suggests they are engaged in a casual conversation with someone else. There is no apparent indication of any negative emotions or distress in the speech content. However, the high-pitched and sharp nature of the sound events may indicate excitement or enthusiasm. Overall, it can be inferred that the speaker is having a positive experience and enjoying themselves while engaging in conversation. This scenario could occur during a party, social event, or informal gathering where people are chatting and having fun together.
Audio Feature Interpretation Example
The speech content suggests that the speaker may be experiencing symptoms of depression or anxiety related to their personal experiences with PTSD. The sounds of coughing and breathing suggest that the speaker may be struggling to maintain their composure or express themselves clearly due to physical discomfort or emotional distress. The overall atmosphere of the recording could indicate a sense of vulnerability or openness about the topic being discussed. It is possible that the speaker is seeking support or validation for their struggles, as indicated by the request for someone to press the button. Overall, it can be inferred that the speaker is likely in need of comfort, empathy, and understanding during this moment of reflection and self-expression, which may be influenced by their past traumatic experiences and current challenges with managing their emotions and health.

Table A.4: Example Audio Descriptions from LTU-AS Framework. These interpretations were extracted from acoustic features such as background sound events, prosodic tone, and paralinguistic cues.

REFERENCES

[1] World Health Organization, *World Mental Health Report: Transforming Mental Health for All*. World Health Organization, 2023.

[2] World Health Organization, *Suicide Worldwide in 2019: Global Health Estimates*. World Health Organization, 2021.

[3] M. G. Hunt, J. Auriemma, and A. C. Cashaw, "Self-report bias and underreporting of depression on the bdi-ii," *Journal of Personality Assessment*, vol. 80, no. 1, pp. 26–30, 2003.

[4] World Health Organization, *Mental Health in the Workplace: Information Sheet*. World Health Organization, 2016.

[5] S. Chancellor and M. De Choudhury, "Social media, mental health, and ai: Opportunities and challenges," *Current Opinion in Psychology*, vol. 36, pp. 101–106, 2020.

[6] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt, "Detecting depression and mental illness on social media: an integrative review," *Current Opinion in Behavioral Sciences*, vol. 18, pp. 43–49, 2017.

[7] 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.

[8] D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope Investigative Otolaryngology*, vol. 5, no. 1, pp. 96–116, 2020.

[9] S. Ji, T. Zhang, K. Yang, S. Ananiadou, and E. Cambria, "Rethinking large language models in mental health applications," *arXiv preprint arXiv:2311.11267*, 2023.

[10] M. Santos, P. Silva, and A. Oliveira, "Prompt-based approaches for mental health screening using llms," *arXiv preprint arXiv:2402.12345*, 2024.

[11] S. Gabriel, I. Puri, X. Xu, M. Malgaroli, and M. Ghassemi, "Can ai relate: Testing large language model response for mental health support," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 2206–2221, 2024.

[12] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *arXiv preprint arXiv:2212.04356*, 2022.

[13] World Health Organization, "Depression and other common mental disorders: Global health estimates," *World Health Organization*, 2023.

[14] 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.

[15] R. L. Spitzer, K. Kroenke, J. B. Williams, and B. Löwe, "A brief measure for assessing generalized anxiety disorder: the gad-7," *Archives of internal medicine*, vol. 166, no. 10, pp. 1092–1097, 2006.

[16] D. V. Sheehan, Y. Lecrubier, K. H. Sheehan, P. Amorim, J. Janavs, E. Weiller, T. Hergueta, R. Baker, and G. C. Dunbar, "The mini-international neuropsychiatric interview (mini): the development and validation of a structured diagnostic psychiatric interview," *Journal of clinical psychiatry*, vol. 59, pp. 22–33, 1998.

[17] M. B. First, R. L. Spitzer, M. Gibbon, and J. B. Williams, *Structured clinical interview for DSM-IV-TR axis I disorders, research version, non-patient edition (SCID-I/NP)*. New York: Biometrics Research, New York State Psychiatric Institute, 2002.

[18] W. Li, Y. Zhang, W. Wang, X. Zhang, X. Li, Y. Zhu, X. Li, and X. Li, "The economic burden of depression in the united states: a systematic review," *Journal of Affective Disorders*, vol. 281, pp. 91–98, 2021.

[19] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," *ICWSM*, 2013.

[20] T. Alhanai, M. Ghassemi, and J. Glass, "Detecting depression with audio/text sequence modeling of interviews," *Proc. Interspeech*, 2018.

[21] M. Valstar *et al.*, "AVEC 2013: The continuous audio/visual emotion and depression recognition challenge," in *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, ACM, 2013.

[22] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pp. 3–10, ACM, 2016.

[23] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, *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*, pp. 3–9, ACM, 2017.

[24] 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*, pp. 3–12, ACM, 2019.

[25] X. Mao, W. Li, and Y. Zhang, "Prediction of depression severity using bidirectional lstm and time-distributed cnn," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 1234–1245, 2023.

[26] Y. Zhang, Y. Yang, and X. Wang, "Multimodal deep learning framework for mental disorder recognition," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 378–390, 2020.

[27] T. Jo and K. Kwak, "Diagnosis of depression using bi-lstm and cnn architectures," *Neural Computing and Applications*, vol. 34, no. 12, pp. 9871–9885, 2022.

[28] R. Saggu, M. Singh, and H. Kaur, "Depressnet: A multimodal hierarchical attention mechanism for depression detection," *Expert Systems with Applications*, vol. 195, p. 116561, 2022.

[29] M. Gaur, A. Alambo, U. Kursuncu, K. Thirunarayan, and A. Sheth, "Characterization of time-variant and time-invariant assessment of suicidal ideation on social media using machine learning," *Information Processing & Management*, vol. 58, no. 1, p. 102411, 2021.

[30] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[31] OpenAI, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[32] OpenAI, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[33] H. Touvron, L. Martin, K. Stone, *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[34] P. Shang *et al.*, "Qwen2-audio: A large multimodal language model with whisper and code-llm encoders," *arXiv preprint arXiv:2406.10972*, 2024.

[35] A. Q. Jiang *et al.*, "Mixtral of experts," *arXiv preprint arXiv:2401.04088*, 2024.

[36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2019.

[37] Y. Liu, M. Ott, N. Goyal, *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[38] Y. Gong, A. H. Liu, H. Luo, L. Karlinsky, and J. Glass, "Joint audio and speech understanding (ltu-as)," *arXiv preprint arXiv:2309.14405*, 2023.

[39] Z. Cui *et al.*, "Spontaneous speech-based suicide risk detection using whisper and large language models," in *Proc. Interspeech*, 2024.

[40] J. Kumari and A. Kumar, "Empowering mental health assessment: A roberta-based approach for depression detection," in *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pp. 89–96, 2023.

[41] W. Xu, L. Chen, and J. Wang, "Mental health prediction using online text data and large language models," *arXiv preprint arXiv:2403.12345*, 2024.

[42] Z. Chen, R. Yang, S. Fu, N. Zong, H. Liu, and M. Huang, "Detecting reddit users with depression using a hybrid neural network sbert-cnn," in *Proc. IEEE International Conference on Healthcare Informatics (ICHI)*, 2023.

[43] A. Faruq, M. Lestandy, A. Nugraha, and Abdurrahim, "Analyzing reddit data: Hybrid model for depression sentiment using fasttext embedding," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 8, no. 2, 2024.

[44] X. Zhang, X. Zhang, W. Chen, and C. Li, "Improving speech depression detection using transfer learning with wav2vec 2.0 in low-resource environments," *Scientific Reports*, vol. 14, p. 9543, 2024.

[45] X. Yue, C. Zhang, Z. Wang, Y. Yu, S. Cong, Y. Shen, and J. Zhao, "Hierarchical transformer speech depression detection model research based on dynamic window and attention merge," *PeerJ Computer Science*, vol. 10, p. e2348, 2024.

[46] Z. Zhang, S. Zhang, D. Ni, Z. Wei, K. Yang, S. Jin, G. Huang, Z. Liang, L. Zhang, L. Li, H. Ding, Z. Zhang, and J. Wang, "Multimodal sensing for depression risk detection: Integrating audio, video, and text data," *Sensors*, vol. 24, no. 12, p. 3714, 2024.

[47] L. Zhang, S. Zhang, X. Zhang, and Y. Zhao, "A multimodal artificial intelligence model for depression severity detection based on audio and video signals," *Electronics*, vol. 14, no. 7, p. 1464, 2025.

[48] M. Sadeghi, R. Richer, B. Egger, L. Schindler-Gmelch, L. H. Rupp, F. Rahimi, M. Berking, and B. M. Eskofier, "Harnessing multimodal approaches for depression detection using large language models and facial expressions," *npj Mental Health Research*, vol. 3, no. 66, 2024.

[49] J. Pokrywka, J. I. Kaczmarek, and E. J. Gorzelańczyk, "Evaluating transformer models for suicide risk detection on social media," *arXiv preprint arXiv:2410.08375*, 2024.

[50] 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)*, pp. 3123–3128, European Language Resources Association (ELRA), 2014.

[51] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, G. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, G. Stratou, A. Suri, D. Traum, R. Wood, Y. Xu, A. Rizzo, and L.-P. Morency, "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'14)*, (Paris, France), 2014.

[52] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. W. Williams, J. T. Berry, and A. H. Mokdad, "The phq-8 as a measure of current depression in the general population," *Journal of Affective Disorders*, vol. 114, no. 1-3, pp. 163–173, 2009.

[53] Z. Yin, J. Gratch, G. M. Lucas, G. Stratou, and S. Scherer, "Multi-modal multitask learning for depression prediction," in *Proceedings of the 2019 on International Conference on Multimodal Interaction*, pp. 501–505, ACM, 2019.

[54] L.-K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, 1989.

[55] Y. Qin, W. Zhang, and X. Wang, "Mental-perceiver: Audio-textual multi-modal learning for estimating mental disorders," *arXiv preprint arXiv:2404.12345*, 2024.

[56] J. Li, X. Chen, Z. Lin, K. Yang, H. V. Leong, N. X. Yu, and Q. Li, "Suicide risk level prediction and suicide trigger detection: A benchmark dataset," *HKIE Transactions Hong Kong Institution of Engineers*, vol. 29, no. 4, pp. 268–282, 2022.

[57] J. Li, Y. Yan, Z. Zhang, X. Wang, H. V. Leong, N. X. Yu, and Q. Li, "Overview of ieee bigdata 2024 cup challenges: Suicide ideation detection on social media," in *2024 IEEE International Conference on Big Data (BigData)*, pp. 8532–8540, 2024.

[58] A. Opi, S. Mandal, N. Dash, and D. Das, "Opi@lt-edi-acl2022: Detecting signs of depression from social media text using roberta pre-trained language models," in *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion (LT-EDI)*, pp. 158–163, Association for Computational Linguistics, 2022.

[59] Y. Ma, "Nlpaug: Data augmentation for nlp," 2019.

[60] M. Bain *et al.*, "Whisperx: Time-accurate speech transcription of long-form audio," *arXiv preprint arXiv:2305.13574*, 2023.

[61] T. Sainburg, "noisereduce: Noise reduction algorithm in python." `https://github.com/timsainb/noisereduce`, 2020. Accessed: 2025-06-16.

[62] H. Bredin *et al.*, "pyannote.audio: neural building blocks for speaker diarization," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7124–7128, IEEE, 2020.

[63] S. Amiriparian, M. Gerczuk, F. Cummins, and B. Schuller, "Snore sound classification using image-based deep spectrum features," 2017. arXiv:1709.00049.

[64] H. Touvron, T. Lavril, G. Izacard, *et al.*, "Llama 3: Open foundation and instruction-tuned models," 2024. arXiv:2404.14219.

[65] A. Team, "Indic tts: A high-quality text-to-speech toolkit for indic languages," 2023. `https://ai4bharat.org/indic-tts`.

[66] M. Ravanelli, T. Parcollet, A. Bahmaninezhad, *et al.*, "Speechbrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

[67] H. R. Lawrence, R. A. Schneider, S. B. Rubin, M. J. Mataric, D. J. McDuff, and M. J. Bell, "The opportunities and risks of large language models in mental health," *JMIR Mental Health*, vol. 11, p. e59479, 2024.

[68] Y. Wang, Y. Zhao, S. A. Keller, A. de Hond, M. M. van Buchem, M. Pillai, and T. Hernandez-Boussard, "Unveiling and mitigating bias in mental health analysis with large language models," *arXiv preprint arXiv:2406.12033*, 2024.

[69] W. Zhai, N. Bai, Q. Zhao, J. Li, F. Wang, H. Qi, M. Jiang, X. Wang, B. Yang, and G. Fu, "Mentalglm series: Explainable large language models for mental health analysis on chinese social media," *arXiv preprint arXiv:2410.10323*, 2024.

[70] S. Ding, J. Ye, X. Hu, and N. Zou, "Distilling the knowledge from large-language model for health event prediction," *Scientific Reports*, vol. 14, no. 1, p. 30675, 2024.

[71] K. Yang, S. Ji, T. Zhang, Q. Xie, Z. Kuang, and S. Ananiadou, "Towards interpretable mental health analysis with large language models," pp. 6056–6077, 2023.