A Multi-Task LLM Framework for Multimodal Speech-Based Mental Health Prediction

Mai Ali[1], Christopher Lucasius[1*], Tanmay P. Patel[1*], Madison Aitken[2,3], Jacob Vorstman[1,4], Peter Szatmari[1,2,4]
Marco Battaglia[1], Deepa Kundur[1]

***Abstract*—Mental health disorders are often comorbid, highlighting the need for predictive models that can address multiple outcomes simultaneously. Multi-task learning (MTL) provides a principled approach to jointly model related conditions, enabling shared representations that improve robustness and reduce reliance on large disorder-specific datasets. In this work, we present a trimodal speech-based framework that integrates text transcriptions, acoustic landmarks, and vocal biomarkers within a large language model (LLM)-driven architecture. Beyond static assessments, we introduce a longitudinal modeling strategy that captures temporal dynamics across repeated clinical interactions, offering deeper insights into symptom progression and relapse risk. Our MTL design simultaneously predicts depression relapse, suicidal ideation, and sleep disturbances, reflecting the comorbid nature of adolescent mental health. Evaluated on the Depression Early Warning (DEW) dataset, the proposed longitudinal trimodal MTL model achieves a balanced accuracy of 70.8%, outperforming unimodal, single-task, and non-longitudinal baselines. These results demonstrate the promise of combining MTL with longitudinal monitoring for scalable, noninvasive prediction of adolescent mental health outcomes.**

***Index Terms*—multi-task learning; multimodal speech analysis; large language models (LLM); depression prediction; longitudinal modeling; digital phenotyping; natural language processing, digital health.**

I. Introduction

Depression, suicidal ideation, and sleep disturbances are prevalent and interconnected mental health issues among adolescents. Major depressive disorder affects 8–12% of adolescents globally, with suicide being the second leading cause of death in individuals aged 15–24 [1]. Sleep disturbances, affecting up to 50% of adolescents, significantly increase the risk of both depression and suicidality [2], with each hour of sleep loss raising suicidal thoughts by 11% [3]. These conditions are bidirectionally linked, suggesting that integrated screening could enhance early detection.

Speech offers a non-invasive avenue for mental health assessment, as individuals with depression or suicidal ideation exhibit distinct linguistic and paralinguistic features [4]. Recent advances in large language models (LLMs) enable automatic detection of these markers. Importantly, Multi-task learning enables the simultaneous modeling of multiple mental health conditions, facilitating shared representation learning that enhances generalization performance and reduces the need for large condition-specific labeled datasets.

Despite advances in LLM-based depression detection, gaps remain. We address them with the following contributions:

1) A comprehensive multimedia framework for depression relapse detection that integrates three speech-derived modalities: speech transcriptions, acoustic landmarks [5], and vocal biomarkers. This approach provides a holistic understanding of speech-based depression indicators.
2) A longitudinal analysis framework that tracks changes across clinical interactions by treating the interactions as an LLM 'conversation'. This type of analysis is not novel, but we are the first to apply it to multimodal speech and text-based mental health analysis.
3) A multi-task learning architecture that extends our trimodal approach beyond depression relapse detection to related clinical assessments, maximizing the utility of data collected across multiple related domains.

We assess these contributions relative to current state-of-the-art methods on the Depression Early Warning (DEW) dataset, which is described in Section III-A.

II. Related Works

The potential of machine learning for predicting mental health conditions is well-established. Prior studies explored modalities like actigraphy, sleep patterns, ecological momentary assessments, facial expressions, and speech characteristics [6]. Advancements in LLMs enhanced their ability to process long-form transcripts and infer underlying cognitive states. These models can extract clinically relevant features from multimodal data for scalable, accurate mental health diagnostics [7].

A. Text- and Speech Semantics-Based Prediction

LLM-based depression detection using text or transcribed speech data is a rich field. Xu et al. [8] benchmarked widespread general-purpose LLMs against mental health classification tasks, finding limited potential in zero-shot and few-shot regimes. Their work introduced two new LLMs—Mental-Alpaca and Mental-FLAN-T5—that were instruction fine-tuned for multi-task mental health classification and outperformed larger mainstream models [8].

Another active research area lies in leveraging the conversational power of LLMs to understand patient histories by using longitudinal methods. Whereas previous works such as [9] employ unimodal methods, we use a multimodal framework by treating each clinical assessment as an episode in a larger LLM interaction (see Section IV-D).

B. Vocal Biomarker-Based Prediction

Several studies explore the link between vocal biomarkers (e.g., fundamental frequency, mel-frequency cepstrum coefficients (MFCC)) and mental disorders. Tasnim et al. introduce a depression detection dataset using hand-curated speech features,

[1] University of Toronto, Canada.
[2] Centre for Addiction and Mental Health, Toronto, Canada.
[3] York University, Toronto, Canada.
[4]The Hospital for Sick Children, Toronto, Canada.
corresponding author: maia.ali@mail.utoronto.ca
[*]These authors contributed equally to this work.

including intensity, MFCC 0-12, zero-crossing rate, and fundamental frequency informed by clinical expertise [10]. Our work integrates these biomarkers with additional modalities suitable for LLM analyses, as detailed in Section IV-A3.

C. Acoustic Landmark-Based Prediction

While transcripts are natural inputs for language models, speech contains rich multimedia features relevant to depression relapse detection. Zhang et al. [5] extract acoustic landmarks—discrete symbols representing linguistic and pronunciation patterns—that complement raw transcripts. Their two-stage approach fine-tunes an LLM with Low-Rank Adaptation (LoRA) matrices [11] to encode landmarks and applies prompt (P)-tuning with an attached classifier for depression prediction, achieving state-of-the-art results. They omit features from speech waveforms (e.g., vocal biomarkers) which are less compliant with LLM-based methods, but central to our model.

D. Multi-Task Learning in the Context of Mental Heath

Multi-task learning (MTL) is a machine learning paradigm in which a single model is trained to perform multiple related tasks simultaneously by sharing common representations. This approach allows the model to learn underlying patterns that are shared among tasks, often leading to better generalization and robustness compared to models trained on individual tasks [12]. Benton et al. demonstrated the effectiveness of MTL in predicting mental health conditions from social media text; they showed that combining demographic attributes and mental states in an MTL framework outperformed single-task models [13].

III. Experimental Setup

In this section, we present the Depression Early Warning (DEW) dataset used in this study. We describe the modalities, the data acquisition protocol, and the pre-processing steps. We then outline task label assignments for depression relapse, suicidal ideation, and sleep disturbances. Lastly, we introduce the experimental design.

A. Dataset and Task Labels

The Depression Early Warning (DEW) dataset, collected at CAMH in Toronto, contains speech recordings from adolescents aged 12–21 with a clinical history of MDD. The sample is predominantly female (70%) and White (50%), reflecting demographic trends commonly reported in adolescent depression research. Each participant was scheduled for up to eight follow-up visits over two years, spaced three to four months apart, where semi-structured interviews were conducted to capture naturalistic speech. For analysis, we consider both cross-sectional predictions, where visits are modeled independently, and longitudinal predictions, which incorporate temporal information across sessions.

The study defines three binary classification tasks: depression relapse (derived from harmonized CDRS and HAM-D scores via equipercentile linking [14]), suicidal ideation (based on PHQ-9 Q9 and MFQ Q19), and sleep disturbances (based on PHQ-9 Q3 and MFQ Q32–33) [15].

B. Experimental Overview

We employ binary classification for the three conditions and use three architectures based on the following sets of modalities: text; text & acoustic landmarks; and text, acoustic landmarks & vocal biomarkers. Each architecture is designed for a particular modality set and is built on two 'LLM bases'—the general-purpose LLaMA-2-7B model [16] and Mental-Alpaca. This enables a comparison across modality sets, and between the two LLMs. We optimize a combined loss function across the three tasks and assess performance with metrics like precision, recall, and balanced accuracy. All subjects are separated into training and test sets to avoid data leakage between subjects.

IV. Methodology

We begin by describing the extraction and tokenization processes for the three modalities leveraged in this work: text, acoustic landmarks, and vocal biomarkers. We then describe the three model architectures and training pipelines.

A. Feature Extraction and Tokenization

The proposed architecture treats the patients' speech data as a trimodal multimedia source composed of text, acoustic landmarks, and audio biomarker features. This section details the extraction of these three components.

1) Text: Text transcripts of the speech data are generated using OpenAI's Whisper Speech Recognition System. While this software is known to be robust [17], we manually scanned the generated transcripts and re-transcribed any incomplete sentences.

2) Landmarks: Acoustic landmarks are extracted as per [5]. Each audio spectrogram is divided into six frequency bands. Energy changes in one or more of these bands are classified under various landmark symbols, like vibration of vocal folds, release or closure of the nasal passage, voiced frication, periodicity, etc. The sequence of landmarks corresponding to a specific speech sample is recorded alongside the text transcriptions for consumption by the proposed architectures.

3) Vocal Biomarkers: Vocal biomarkers are extracted using Python's Librosa library [10]. We divide each speech sample into 500-ms windows and extract their summary statistics. These features encompass spectral characteristics like sound intensity, MFCC, delta-MFCC, pitch, magnitude, and zero-crossing rate (ZCR), along with voicing-related attributes like fundamental frequency (F_0), harmonicity, harmonic-to-noise ratio (HNR), shimmer and jitter, energy, durational features, pauses, fillers, and phonation rate.

B. Baseline A: Text Model with MTL

This baseline employs a P-tuned version of either Mental-Alpaca (see Section II-A), or Meta AI's LLaMA-2-7B [16]. Both models have potential since Mental-Alpaca is fine-tuned to mental health tasks, while LLaMA-2-7B has shown promise on general-purpose prediction. The LLM's predicted embedding is then used for MTL, as per Section IV-E.

C. Baseline B: Text and Acoustic Landmark Pipeline with MTL

This pipeline is heavily influenced by [5]. We replicate their two-stage procedure: hint cross-modal instruction fine-tuning, followed by P-tuning for depression detection [18]. Similar to

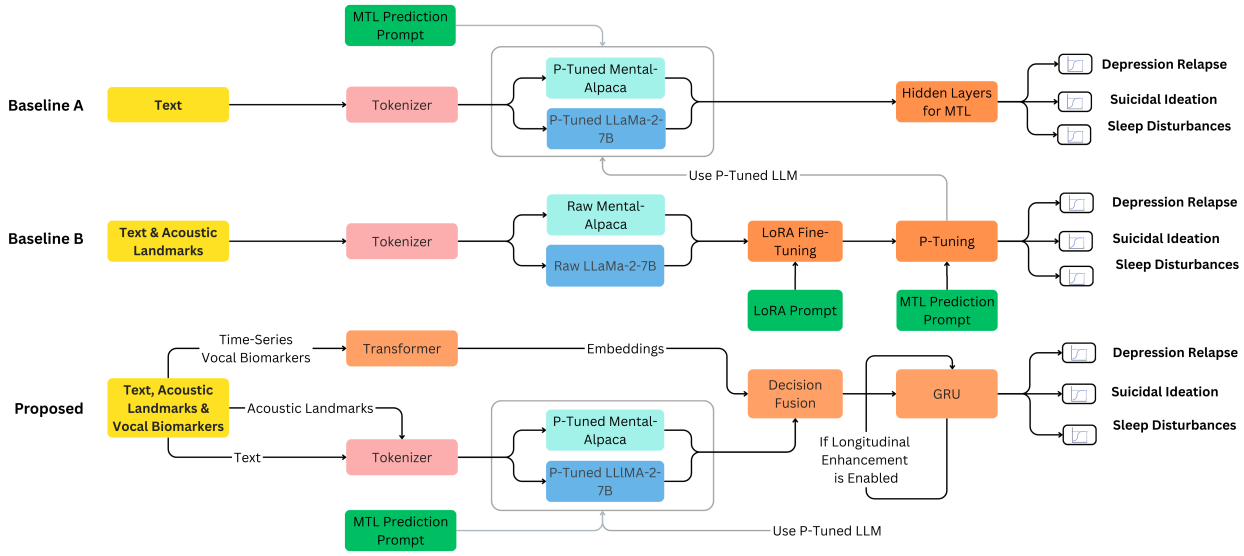

Fig. 1. The two baseline architectures use subsets of the available modalities. Our proposed architecture leverages all available modalities *and* analyses them longitudinally over a patient's history. All architectures use an MTL framework and a common fine-tuning strategy to ensure fair comparisons.

baseline A, we extend the work to support MTL and allow support for both LLaMA-2-7B and Mental-Alpaca.

In cross-modal fine-tuning, the LLM is prompted in a manner that elicits correspondances between acoustic landmarks and its transcript. This allows the LLM to learn the semantics of acoustic landmarks and align the positions of landmarks to text. Since the LLM has billions of parameters, we use the LoRA technique [8] to only train a limited parameter subset. After the LLM is trained to recognize the combined text and landmark data, a different prompt is fed into the LLM to ask it to predict three binary labels: depression, suicidal ideation, and sleep disturbances. Instead of using the usual language model head, a fully connected classification layer is added to the LLM to make these predictions. P-tuning is applied to add trainable prompt embeddings in combination with the original prompt to fine-tune the LLM during training.

D. Proposed Pipeline: Text, Acoustic Landmark, and Vocal Biomarkers for Multi-Task, Longitudinal Analysis

This architecture unifies all three speech-derived modalities into a novel, multimodal depression detection system. It also introduces time-awareness by supporting longitudinal analysis across multiple clinical visits. A depiction of the architecture, along with the baselines, is provided in Fig. 1.

The model uses the result of baseline B to generate final embeddings from the text data and their corresponding acoustic landmarks. Since this model was pre-trained to analyze text and acoustic landmark samples for depression, its weights can remain frozen.

To analyze the vocal biomarkers extracted from the speech samples, we generate contextualized embeddings through a transformer encoder that captures the time-series nature of the biomarkers. The two embeddings are fused at the decision-level with trainable weights which are then used for MTL.

This architecture allows the user to optionally track a hidden latent vector between subsequent patient visits. This vector is propagated from visit to visit through a Gated Recurrent Unit (GRU), introducing a second layer of temporal analysis beyond the inherent time-series nature of speech.

E. MTL Formulation (For All Architectures)

For all three pipelines, the final embeddings pass through three separate heads responsible for one task each. The gradients between heads are not detached, allowing decisions from one task to influence the others and thus enabling us to leverage comorbidity (the is evaluated in Section V-C). Since each task can suffer from class imbalance, we use a weighted binary cross-entropy loss. The loss for a single task t is

$$\mathcal{L}_t = -\left(w_t^+ y_t \log(\hat{y}_t) + (1 - y_t)\log(1 - \hat{y}_t)\right) \tag{1}$$

where y_t is the true label, $\hat{y}_t$ is the predicted probability, and w_t^+ is the weight applied to positive samples to account for task-specific class imbalance. The total loss, then, is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_M + \lambda_{\text{aux}} \left(\mathcal{L}_{A_0} + \mathcal{L}_{A_1}\right) \tag{2}$$

where $\mathcal{L}_M$ is the main task loss, $\mathcal{L}_{A_0}$ and $\mathcal{L}_{A_1}$ are auxiliary task losses, and λ_{aux} controls the weight of auxiliary losses.

V. Results and Analysis

Table I presents the classification performance across all three binary prediction tasks, categorized based on the subset of modalities utilized, the corresponding model architectures, and the base LLM employed. To ensure consistency, $\lambda_{\text{aux}} = 0.25$ is used for all trials. While precision and recall are reported for completeness, balanced accuracy serves as our primary metric since it best captures aggregate performance across both positive and negatives cases.

To ensure a fair study, a Receiver Operating Characteristic (ROC) curve is constructed for each case based on the validation set, and thresholds are selected based on the point on the curve closest to the top left corner. This threshold is then blindly applied to the test set and the resulting metrics are reported to ensure no information leakage between sets.

A. Effect of Modality and Base LLM Architecture

Table I shows an improvement in balanced accuracy as more modalities are incorporated into the Mental-Alpaca-based analysis, highlighting the utility of multimodal approaches. This

TABLE I
MTL RESULTS ACROSS MODALITIES/MODELS; MA=MENTAL-ALPACA, L2=LLAMA-2-7B, P=PRECISION, R=RECALL, BA=BALANCED ACCURACY

Architecture / Modalities	Depression Relapse						Suicidal Ideation						Sleep Disturbances					
	MA			L2			MA			L2			MA			L2		
	P	R	BA	P	R	BA	P	R	BA	P	R	BA	P	R	BA	P	R	BA
Text	0.362	0.607	0.606	0.258	0.537	0.524	0.000	0.000	0.500	0.000	0.000	0.500	0.885	1.000	0.500	0.885	1.000	0.500
Text & Acoustic Landmarks	0.291	0.460	0.518	0.302	0.520	0.533	0.215	0.472	0.524	0.219	0.694	0.542	0.929	0.404	0.583	0.898	0.273	0.518
Proposed: All Modalities	0.402	0.660	0.644	0.185	0.300	0.400	0.727	0.222	**0.601**	0.201	0.778	0.509	1.000	0.099	0.550	0.885	1.000	0.500
Proposed: All Modalities with Longitudinal Enhancement	0.425	0.680	**0.666**	0.271	0.580	0.495	0.235	0.639	0.563	0.286	0.056	0.511	0.924	0.752	**0.638**	0.957	0.280	0.592

is most evident in the Suicidal Ideation task, where precision and recall start at 0 with text alone and increase substantially as additional modalities are incorporated. We also observe a significant improvement when using Mental-Alpaca, compared to LLaMA-2-7B, which agrees with the literature [8].Despite LLaMA-2-7B's size, Mental-Alpaca's domain pretraining yields superior predictive performance, achieving the highest balanced accuracy across all tasks in the trimodal setting.

B. Effect of Longitudinal Analysis

Table I demonstrates that the longitudinal enhancement to the full trimodal pipeline results in improved balanced accuracy for all three tasks with LLaMA-2-7B and two tasks with Mental-Alpaca. This shows that monitoring long-term trajectories leads to improved predictive power, aligning with previous work on unimodal text datasets [9].

Note that for the Suicidal Ideation task with Mental-Alpaca, enabling the longitudinal enhancement causes precision to fall and recall to rise. We hypothesize this occurs because incorporating patient history improves sensitivity to emerging risk patterns, increasing true positives but also introducing more false positives.

C. Utility of MTL

For this experiment, the trimodal Mental-Alpaca-based longitudinal architecture is used, since it was found to perform best in the previous study. Table II demonstrates the effect of increasing the importance (weightage) applied to the two auxiliary tasks on the ability to predict the main task as illustrated in (2). Increasing the weights on the auxiliary tasks leads to improved balanced accuracies for the main task, with the exception of $\lambda_{\text{aux}} = 0.75$. This suggests synergistic effects among the three tasks, best exploited through joint learning.

TABLE II
EFFECT OF AUXILIARY TASK WEIGHTS ON PRIMARY DEPRESSION TASK

Auxiliary Tasks Weights	Primary Task Metrics		
	P	R	BA
0.00	0.354	0.580	0.589
0.25	0.438	0.420	0.608
0.50	0.427	0.700	0.672
0.75	0.468	0.580	0.665
1.00	0.535	0.620	**0.708**

VI. CONCLUSION AND LIMITATIONS

Our results demonstrate that deploying a longitudinal LLM-based model on speech data—treated as a trimodal multimedia source—enhances performance in predicting multiple mental health outcomes. By leveraging speech digital phenotypes, our approach captures rich behavioural markers, leading to improved performance compared to baseline models [18] on our dataset. To further substantiate these findings, we plan to benchmark our pipeline against publicly available datasets that have been used to assess baseline models, including Mental-Alpaca and other LLMs, in the context of comparable tasks. This work can also be extended to explore other LLM architectures, such as Mental-FLAN-T5 and GPT. Given that LLM outputs are heavily influenced by input prompts, conducting a more in-depth analysis of different prompt strategies would provide valuable insights.

One of the limitations of this work is that the dataset's demographic composition is skewed, with a predominance of female and White participants. This imbalance may restrict the generalizability of the results and raise concerns about potential bias in predictive performance for underrepresented racial and gender groups.

REFERENCES

[1] W. Lu, "Adolescent depression: National trends, risk factors, and healthcare disparities," *Am J Health Behav*, 2019.

[2] M. Gradisar et al., "Sleep's role in the development and resolution of adolescent depression," *Nature Reviews Psychology*, 2022.

[3] S. Inkelis et al., "Elevated risk of depression among adolescents presenting with sleep disorders," *J. Clin. Sleep Med*, 2020.

[4] C. M. Corcoran et al., "Language as a biomarker for psychosis: A natural language processing approach," *Schizophrenia Research*, 2020.

[5] X. Zhang et al., "When llms meets acoustic landmarks: An efficient approach to integrate speech into large language models for depression detection," 2024.

[6] C. Lucasius et al., "Prediction of relapse in adolescent depression using fusion of video and speech data."

[7] E. C. Stade et al., "Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation," *npj Mental Health Research*, 2024.

[8] X. Xu et al., "Mental-llm: Leveraging large language models for mental health prediction via online text data," *IMWUT*, 2024.

[9] W. Qin et al., "Read, diagnose and chat: Towards explainable and interactive llms-augmented depression detection in social media," 2023.

[10] T. Mashrura et al., "Depac: a corpus for depression and anxiety detection from speech," 2023.

[11] E. J. Hu et al., "Lora: Low-rank adaptation of large language models," 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

[12] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, pp. 41–75, 1997.

[13] A. Benton et al., "Multitask learning for mental health conditions with limited social media data," *ACL Anthology*, pp. 152–162, 04 2017.

[14] T. Furukawa et al., "Translating the BDI and BDI-II into the HAMD and vice versa with equipercentile linking," *Epidemiol. Psychiatr. Sci.*, 2019.

[15] G. E. Simon et al., "Does response on the phq-9 depression questionnaire predict subsequent suicide attempt or suicide death?" *Psychiatric Services*, 2013.

[16] H. Touvron et al., "Llama: Open and efficient foundation language models," 2023.

[17] A. Radford et al., "Robust speech recognition via large-scale weak supervision," 2022.

[18] X. Liu et al., "GPT Understands, Too," *arXiv:2103.10385 [cs]*, 2021. [Online]. Available: https://arxiv.org/abs/2103.10385