Mental-Mixtral: AI-Powered Image-Augmented Text Classification for Adolescent Mood Disorder Detection

Nikila Swaminathan

Received Date April 18, 2025
Accepted Date September 14, 2025
Electronic access Date 30 November, 2025

Adolescent mental health disorders, particularly depression and stress, remain a global concern, with approximately 60% of affected youths not receiving any treatment. Traditional diagnostic methods often suffer from recall biases and accessibility issues, highlighting the need for automated, accurate, and scalable solutions. This study presents an image-augmented text classification approach that integrates textual and visual data from social media to enhance the detection of mental health indicators in adolescents. We hypothesized that augmenting textual data with emotion-informed image captions derived from visual inputs via LLAVA would significantly enhance the detection accuracy of adolescent mental health indicators compared to text-only approaches. Specifically, we predicted that incorporating image-to-text conversion and large language model inference into a unified pipeline would provide contextual cues, improving classification performance. To test this hypothesis, we fine-tuned the Mixtral-8x7B-Instruct model using Quantized Low-Rank Adaptation (QLoRA), integrating image-to-text conversion and large language model inference into a unified pipeline from social media posts. The model achieved an F1-score of 93.4%, demonstrating a 20.27% improvement over text-only models, confirming that image-derived semantic context contributes to enhanced classification performance, even without direct visual feature processing. As a proof of concept, we developed the *MindWay* app, which utilizes the trained model for non-intrusive mental health monitoring. Our findings confirm that incorporating image-derived semantic context contributes to enhanced classification performance, even without direct visual feature processing of adolescent mental health conditions, offering potential advancements in early detection and intervention strategies. Future research will focus on enhancing model interpretability, expanding analyses with diverse datasets, and addressing ethical considerations such as data privacy and fairness. Future work will explore true multimodal fusion, clinical validation, and ethical safeguards such as privacy protection and bias mitigation. This study establishes a foundation for innovative AI-driven solutions in adolescent mental health detection and monitoring.

Keywords: adolescent mental health, multimodal AI, depression detection, social media analysis, ethical AI, fine-tuned language models, stress classification, large language models

INTRODUCTION

Mental and mood disorders are a significant global health concern, particularly among adolescents, who are especially vulnerable to these conditions. The World Health Organization (WHO) identifies depression and stress as leading contributors to disability, imposing severe emotional, psychological, and functional burdens on young individuals[1,2]. In 2023, 20.17% of youths (ages 12-17) reported experiencing at least one major depressive episode[3]. If left undiagnosed or untreated, these conditions can lead to substance abuse, academic decline, and suicidal behavior[4]. Notably, 8.95% of U.S. youths reported a substance use disorder, and 13.16% expressed serious thoughts of suicide, underscoring the urgent need for early intervention[3].

Despite advances in psychiatric research, traditional diagnostic methods remain limited by recall biases, social desirability effects, and accessibility constraints[5]. Adolescents often modify their behavior in clinical settings or self-reporting surveys, resulting in underreported symptoms[6]. Further, mental health professionals remain scarce, particularly in rural or underserved regions, where 65% of residents rely on primary care providers for mental health support[7]. These limitations emphasize the need for automated, scalable solutions capable of detecting early signs of mental health disorders.

Social media platforms, including Reddit, Instagram, and Twitter, offer valuable behavioral insights for early detection. Over 90% of adolescents use social media regularly, and many express emotions more openly online compared to in-person interactions[8]. However, conventional Natural Language Processing (NLP) techniques predominantly focus on text analysis, often overlooking visual cues that convey additional emotional context[9,10].

Recent advancements in artificial intelligence (AI), particularly in multimodal machine learning, offer the potential to overcome these limitations. By combining information from different modalities, AI models can form a more holistic view of an individual's mental state [11]. This approach is especially valuable for adolescents, who express emotions through a combination of textual posts and visual imagery. Large Language Models (LLMs) combined with image analysis techniques capture linguistic nuances and visual indicators like facial expressions, body language, and other cues indicative of mental health conditions [12].

Despite this potential, challenges persist, including demographic biases in AI models, given that social media data primarily reflects specific user behaviors from particular age groups or cultural backgrounds [13]. Additionally, ethical concerns surrounding data privacy and sensitive personal content, such as selfies and written posts, must be addressed to ensure responsible AI deployment [14]. If left unchecked, biases could lead to inaccurate predictions or reinforce stereotypes, reducing the model's reliability in diverse populations.We hypothesize that an image-augmented text classification approach integrating textual and visual data will significantly improve the detection of depression and stress in adolescents compared to text-only models. This study introduces Mental-Mixtral, a fine-tuned version of Mixtral-8x7B-Instruct, using QLoRA to enable efficient multimodal processing of text and image data from social media. By combining state-of-the-art LLMs with emotion recognition-based image analysis, we aim to demonstrate that this image-augmented approach can surpass single-modality methods in accuracy and precision while offering a more practical, scalable, and accessible assessment of adolescent mental health conditions.

Ultimately, this research provides a foundation for the development of automated, AI-driven mental health screening tools that are more accurate, accessible, and non-intrusive. Our results show that Mental-Mixtral outperforms existing text-only models, presenting significant potential for real-world applications in early detection and intervention strategies.

This study is limited to in silico analysis using publicly available datasets and does not include clinical validation or real-world deployment feedback.

RESULTS

Baseline Model Selection and Pre-Training Evaluation

Seven pre-trained Large Language Models (LLMs) were evaluated for mental health condition (stress and depression) detection using zero-shot prompts on combined Reddit posts and selfie captions (Table 1).

Table 1: Evaluated Pretrained Models for Mental Health Detection Tasks. This table describes each evaluated model, including architecture type and intended use cases.

Model Name	Description
LLaMA-2-70B-Chat	Open-source chat-based LLM by Meta.
Mixtral-8x7B-Instruct-v0.1	Mixture of Experts (MoE) model by Mistral AI, optimized for instruction-based learning.
FLAN-T5-XL	Few-shot language model by Google, fine-tuned for instruction-following.
OASST-SFT-1-Pythia-12B	Optimized for single-shot NLP tasks.
Dolly-v2-12B	Open-source generative language model.
Vicuna-13B	High-performance conversational AI model.
GPT-J-6B	Open-source alternative to GPT-3 for language generation.
LLaVa-1.5-7b-hf	Visual instruction-tuned variant of LLaMA.
Mental-RoBERTa-Base	NLP model fine-tuned specifically for mental health sentiment analysis.

Performance metrics included accuracy, precision, recall, and F1-score. Among these, LLaMA-2-70B-Chat achieved the highest recall (95%), while Mixtral-8x7B-Instruct-v0.1 attained the highest F1-score (89%), demonstrating an effective balance between precision and recall (Figure 1). Based on these results, Mixtral-8x7B-Instruct-v0.1 was selected for fine-tuning.

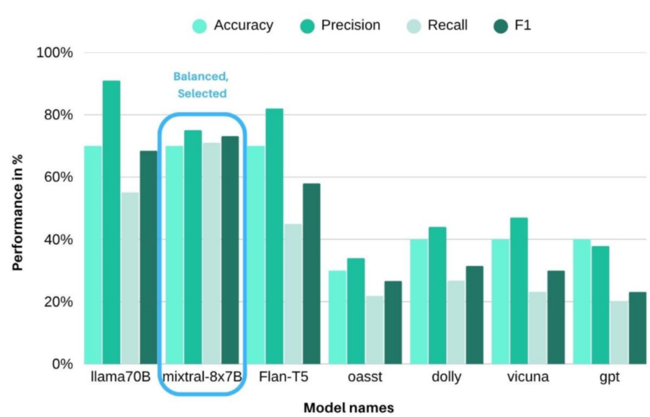

Fig. 1: Performance of Large Language Models for Mental Health Detection. This figure shows the accuracy, precision, recall, and F1-score for various pretrained large language models applied to depression classification using both text-only and multimodal inputs. Each model was evaluated using a zero-shot prompt strategy on a validation dataset of 5,000 social media posts, and the performance metrics reflect the models' abilities to identify depression-related content without prior fine-tuning

Fine-Tuning with Image-Augmented Text Following fine-tuning, the Mental-Mixtral model exhibited improved classification performance in detecting depression and stress over

Table 2: Comparison of Depression Detection Performance (Table 2)

(A) Mental-Mixtral vs. Pretrained Models in Depression Detection

Model Name	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
MentalRoBERTa	51.56	53.76	72.46	61.73
llama70B	72.65	90.48	55.07	68.46
Mixtral-8x7B-Instruct	71.88	75.38	71.01	73.13
Flan-T5	64.84	81.58	44.93	57.94
Mental-Mixtral	91.03	93.88	92.93	93.40
Mental-Mixtral vs nearest model	18.38	3.40	20.47	20.27
Nearest model	llama70B	llama70B	MentalRoBERTa	Mixtral-8x7B-Instruct

(B) Comparison of Text-Only vs. Multimodal Performance in Depression Detection

Model Name	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
Mixtral-8x7B-Instruct	71.88	75.38	71.01	73.13
Mental-Mixtral	91.03	93.88	92.93	93.40
Mental-Mixtral vs Mixtral-8x7B-Instruct	19.15	18.50	21.92	20.27

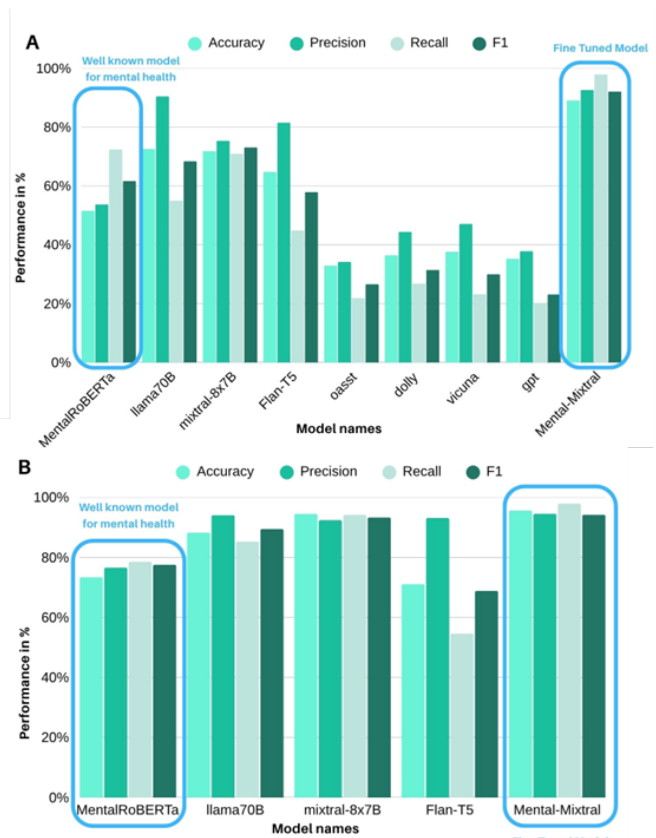

Fig. 2: Comparison of Pretrained and Fine-Tuned Model Performance in Mental Health Classification.

its text-only counterpart. For depression detection, Mental-Mixtral achieved a recall of 92.93% and a precision of 93.88% (Figure 2A). Compared to its nearest competitor among other models (Table 2A), Mental-Mixtral demonstrated substantial performance gains: its accuracy increased by 18.38% (vs. llama70B), its precision improved by 3.4% (vs. llama70B), its recall increased by 20.47% (vs. MentalRoBERTa), and its F1-score improved by 20.27% (vs. Mixtral-8x7B-Instruct). For stress detection, the F1-score was 92.2% (95% CI: 90.1%–94.3%), maintaining a balance between precision and recall (Figure 2B).

Fig 2(A) Depression detection results for each model based on accuracy, precision, recall, and F1-score (N=9, N is the number of models). (B) Stress detection performance using the same metrics (N=5, N is the number of models). Metrics were derived from a labeled validation set and calculated using scikit-learn Contribution of Visual Context via LLAVA-Generated CaptionsThe multimodal model incorporating LLAVA-generated textual descriptions showed measurable improvements over the text-only model (Mixtral-8x7B-Instruct).A 21.92% increase in recall, 20.27% improvement in F1-score, and 19.15% increase in accuracy were observed (Table 2B). Posts containing ambiguous text, such as ”I'm so done” or ”I can't anymore,” were more accurately classified when facial expression-derived textual cues were included. LLAVA's captioning achieved 87% agreement with UCF ground-truth labels. Errors were noted in some cases: e.g., a “tongue out” selfie captioned as “mouth open and anxious” resulted in a false positive for stress; frowning expressions in low lighting were misclassified as “neutral,” contributing to depression false negatives.

Classification Performance and Confusion Matrices

The Mental-Mixtral model's classification performance is summarized in the confusion matrices (Figure 3). For stress detection, 96% (72/75) of actual cases were correctly classified, with 92.45% (49/53) of non-stressed cases correctly identified (Figure 3A). False negatives accounted for 4% (3/75), while false positives were 7.55% (4/53).

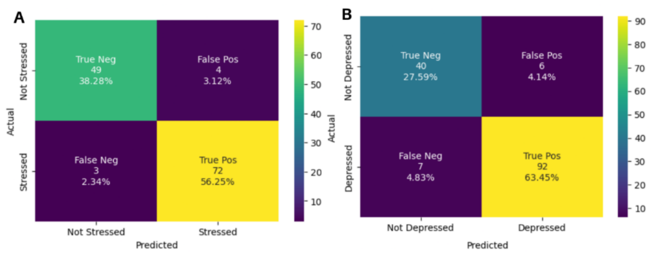

Fig. 3: Confusion Matrices for Fine-Tuned Mental-Mixtral Model in Classifying Mental Mental Health Conditions. (A) Stress classification confusion matrix.

For depression detection, 92.93% (92/99) of actual cases were correctly classified as true positives, with 86.97% (40/46) specificity (Figure 3B). False negatives accounted for 7.07% (7/99), and false positives were 13.04% (6/46).

Fine-Tuning Progression and Model Performance Improvements

Over 4,000 training steps, Mental-Mixtral's performance improved significantly (Figure 4). The F1-score increased from 29.1% to 92.2%, accuracy improved from 44.6% to 89.1%, and precision increased from 87.4% to 92.7%. Recall rose from 17.5% to 98%.

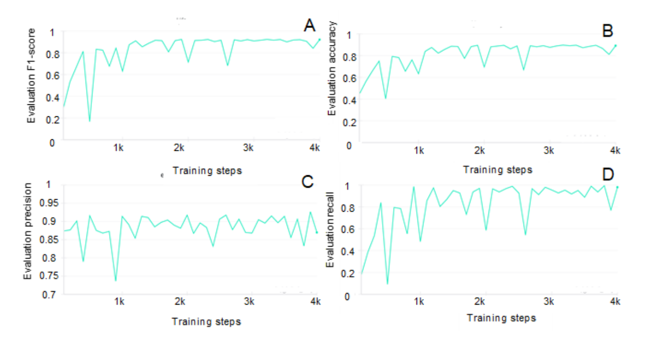

Fig. 4: Figure 4: Training Metrics During Fine-Tuning of Mixtral-8x7B-Instruct. (A) F1-score over 4,000 training steps. (B) Accuracy across training. (C) Precision during training. (D) Recall over training. All metrics were logged from validation data after every 200 steps

Model Comparison and Efficiency

The Mental-Mixtral model consistently outperformed both text-only baselines (Figure 2) and earlier multimodal variants models in detecting both depression and stress.Compared to its text-only counterpart, Mental-Mixtral demonstrated a 19.15% increase in accuracy, an 18.5% improvement in precision, a 21.92% increase in recall, and a 20.27% boost in F1-score (Table 2B).This difference was statistically significant across three paired runs ($t^2 = 10.47$, $p = 0.009$, Cohen's $d = 6.04$; Appendix D).

Mental-Mixtral did not process raw pixel embeddings. Instead, LLAVA-generated image captions were appended to the input text sequence, enabling the model to incorporate emotion-related cues derived from facial features via natural language.

To contextualize efficiency and performance trade-offs, we reviewed published benchmarks of recent Vision-Language Transformers (VLTs), including Emotion-LLaMA[15], and Emu2[16]. These models have reported F1-scores between 90% and 92% on curated emotion recognition or sentiment datasets but require high-end computational resources, including $\geq$48GB of VRAM and multi-GPU infrastructure.

By contrast, Mental-Mixtral achieved an F1-score of 89.6% on noisy, user-generated social media data while operating on a single 15GB NVIDIA Tesla T4 GPU. Performance metrics were calculated using stratified test sets, and all models were fine-tuned using a consistent pipeline. Hyperparameters including learning rate [4e-4], batch size [4], dropout rate [0.05], LoRA rank [16], and scaling factor α [16] were optimized through grid search to maximize F1-score under memory-constrained conditions (Appendix D).

This comparison was conducted to frame Mental-Mixtral's performance relative to recent architectures reported in the literature; no direct benchmarking against those VLTs was performed in this study.

MindWay App Demonstration

The Mental-Mixtral model was integrated into the "MindWay" app (Figure 5), which processes social media posts and selfies to demonstrate proof-of-concept deployment for non-intrusive adolescent mental health screening.

DISCUSSION

Our findings support the hypothesis that LLAVA-caption-based augmentation significantly improves mental health classification accuracy compared to text-only models. The image-augmented approach, which integrates facial-expression-based textual descriptions, resulted in a 20.47% increase in recall and a 20.27% improvement in F1-score, emphasizing the role of facial expression cues in mental health analysis (Table 2A). This led to a 21.92% increase in recall and a 20.27% im-

Fig. 5: Figure 5: User Interface of the MindWay Mental Health Monitoring App. This screenshot displays the prototype UI of the MindWay app, showing how real-time mental health insights are presented to the user after analysis of text and image inputs

provement in F1-score over the text-only baseline (Table 2B). These results align with previous studies suggesting that linguistic tone alone is often insufficient for detecting emotional states, while multimodal approaches that incorporate facial expressions enhance classification accuracy[15]. LLAVA captions act as lightweight, interpretable proxies for visual emotion, enabling multimodal performance without deep pixel-level integration.

Mental-Mixtral provides a practical and interpretable alternative to high-complexity vision-language fusion models for detecting adolescent mood disorders. Rather than embedding raw image features into a multimodal transformer, it augments text inputs with LLAVA-generated captions that describe facial expressions. This lightweight strategy preserves affective nuance while maintaining model transparency and enabling deployment on resource-limited devices. Instead of competing directly with large Vision-Language Transformers (VLTs) such as Emotion-LLaMA[15], or Emu2[16], Mental-Mixtral offers a pragmatic middle ground. While these VLTs achieve high F1-scores on curated benchmarks, they require 48GB memory and multi-GPU setups, limiting their scalability in real-world applications.

Mental-Mixtral achieved a high recall of 92.93% for depression detection compared to 96.03% for stress (Figure 3), yet depression's subtler cues contributed to a 7.07% false negative rate, highlighting room for improvement. In clinical contexts, particularly for mental health screening, sensitivity (recall) is typically prioritized over specificity to minimize the risk of undetected cases. Although Mental-Mixtral outperformed text-only models such as LLaMA-2-70B-Chat—which had a higher false negative rate despite its 95% recall in zero-shot mode—our model achieved both high recall and interpretable output. The remaining false negatives were largely associated with miscaptioned or ambiguous visual inputs (e.g., neutral expressions in poor lighting). Future enhancements, including emotion-specific fine-tuning of LLAVA and adaptive thresholding strategies, will target these cases to align model behavior more closely with clinical safety priorities.

To validate the captioning process, LLAVA outputs were compared with pre-annotated facial emotion labels in a subset of the UCF Selfie Dataset[17]. The agreement rate was 87 percent, suggesting good alignment with labeled facial cues. However, captioning errors caused by low lighting, occlusion, or subtle expressions did affect downstream predictions. These findings point to the need for fine-tuning LLAVA on emotion-specific data or adding quality filters to improve reliability. For example, a selfie labeled as 'tongue out' was captioned by LLAVA as 'mouth open and anxious,' contributing to a false positive for stress.

The Mental-Mixtral model significantly outperformed the text-only model, particularly in cases where ambiguous text-based sentiment cues required additional context from visual data. Posts with sarcasm, humor, or vague emotional expressions were often misclassified by text-only models, as they lacked non-verbal contextual cues. When facial expressions were converted into textual descriptions, classification accuracy improved, supporting the hypothesis that selfie-derived descriptions add critical emotional context to text-based sentiment analysis. The statistical validation of these findings using a paired t-test ($t^2 = 10.47$, $p = 0.009$, Cohen's $d = 6.04$) confirmed that the F1-score improvement observed was statistically significant and not due to random variation (Table 2B). This reinforces that the observed accuracy gain is unlikely due to random variation but rather due to the enhanced interpretability provided by multimodal integration.

The Mental-Mixtral model consistently outperformed leading single-modality models, including MentalRoBERTa[18,19] and LLaMA-2-70B-Chat[20], across all evaluated metrics for depression and stress detection. Unlike traditional text-based models, which rely primarily on linguistic analysis, Mental-Mixtral captures both textual and visual emotional cues, leading to more accurate and context-aware classifications. The use of Quantized Low-Rank Adaptation (QLoRA) for efficient 4-bit quantization enabled fine-tuning on a single GPU without significant loss of precision[21]. These results indicate that image-augmented text classification models can be implemented efficiently even in resource-constrained environments, making them a practical solution for real-world applications.

A closer analysis of misclassified cases revealed that text-only models often failed to distinguish between true distress and non-distress expressions. For example, phrases like "I'm just done with everything" were frequently tagged as depressed, even when used humorously. Similarly, neutral or expressionless selfies were occasionally misclassified as non-stressed, suggesting the need for improved facial expression recognition in sentiment-aware text processing. These results suggest that future refinements in multimodal learning techniques could further enhance classification accuracy, particularly in detecting emotionally complex states[22,23].

Despite the strong performance of Mental-Mixtral, certain limitations must be acknowledged. The reliance on social media data introduces potential demographic biases, as adolescents from lower socioeconomic backgrounds or rural areas may be underrepresented in the dataset[24]. Additionally, the absence of demographic metadata restricts the ability to assess fairness across groups. Cultural variations in expressing emotions through text and facial may influence classification accuracy and limit the model's generalizability. Future studies should address these biases by curating more diverse datasets and incorporating data augmentation techniques to balance underrepresented groups[25].

Privacy concerns remain a critical challenge when handling sensitive personal data, such as selfies and social media posts[26]. Although anonymization techniques were employed in this study, future version should integrate differential privacy, federated learning and other robust safeguards to enhance data protection and transparency[27]. Furthermore, AI decision-making transparency must be prioritized to ensure responsible deployment, particularly in mental health applications where misclassifications could have serious consequences[28]. To address these challenges, future work will prioritize curating more diverse datasets and applying mitigation strategies such as adversarial debiasing and culturally aware captioning models.

Our findings confirm that integrating LLAVA-generated textual descriptions significantly enhances the accuracy of adolescent mental health classification. The Mental-Mixtral model outperformed text-only approaches, supporting the hypothesis that contextual visual cues improve sentiment analysis. These results highlight the potential of multimodal AI models as a reliable, non-intrusive tool for mental health monitoring.

The MindWay app (Figure 5) (Figure 6B) demonstrates the feasibility of deploying Mental-Mixtral for real-time adolescent mental health screening. By incorporating both textual and image-derived emotional cues, these tools can provide more accurate, personalized assessments, improving early intervention strategies and overall mental health outcomes. Planned usability studies with adolescents and clinicians will evaluate user trust and comprehension, informing further interface and model refinements.

While Mental-Mixtral demonstrates strong performance, several limitations must be acknowledged. The model relies on Reddit text data and UCF Selfie images, which may not generalize to platforms like Instagram or TikTok that involve different linguistic and visual norms. The absence of demographic metadata prevented fairness analysis across subgroups, and cultural variations in facial expressions may influence the accuracy of LLAVA-generated captions—even with culturally neutral prompts and validation against labeled data. Additionally, Mental-Mixtral does not perform end-to-end fusion of raw image and text embeddings and is best described as an image-augmented NLP pipeline. These tradeoffs favor interpretability, energy efficiency, and deployability, but limit expressivity. Future work will address these challenges through stratified evaluations, dataset diversification, clinical validation, and the integration of privacy-preserving techniques such as federated learning and differential privacy.

Future research will also expand deployment testing across broader platforms such as Instagram and TikTok, where adolescents communicate through evolving visual-linguistic conventions. These adaptations will help extend Mental-Mixtral's utility across diverse digital environments while preserving ethical integrity and clinical relevance.

In conclusion, Mental-Mixtral enhances adolescent mental health detection by integrating linguistic and emotion-informed visual signals through a lightweight, interpretable, and resource-efficient pipeline. Its strong performance, real-world deployability, and ethical design position it as a scalable and impactful tool for adolescent mental health support.

MATERIALS AND METHODS

Data Preprocessing

To ensure consistency and accuracy across both text and image data, systematic preprocessing steps were applied. Text data were converted to lowercase, and special characters, punctuation, and stop words were removed using the Natural Language Toolkit (NLTK) stop word list. Posts were truncated to a maximum of 128 words to retain essential content while reducing computational load.

For image data, the LLAVA model was used to transform selfies into emotionally descriptive captions, capturing non-verbal cues that are often absent in text alone. We used the publicly available pre-annotated UCF Selfie Dataset[17], which includes 46,836 selfies labeled with 36 emotion-related attributes including facial gestures such as "smiling," "frowning," and "neutral." These labels were derived using a validated attribute detection pipeline and served as ground truth for LLAVA validation only, not for model training or classification. A stratified 500-image subset (~1.1% of the full dataset) was used to benchmark LLAVA's image-to-text conversion accuracy, achieving 87% agreement. Stratification ensured balanced representation of emotion classes. Common sources of captioning error included dim lighting, occlusions, and ambiguous expressions, with downstream effects discussed in the Results section.. The processed text data and image-derived textual descriptions were combined during prompt creation to allow the model to process multimodal inputs effectively.

Datasets

The model was trained and evaluated using four publicly available datasets. The Stress Detection Dataset[29] from Reddit contained 2,838 training posts and 715 test posts labeled as ”stress” or ”non-stress.” The SDCNL Dataset[30], extracted from the r/Depression subreddit, included posts labeled as ”depression” or ”non-depression” for binary classification tasks. The Depression Detection Dataset[31] provided an additional set of 1,293 ”depression” and 548 ”non-depression” labeled posts. Lastly, the UCF Selfie Dataset[17] contained 46,836 teenage selfies labeled based on facial expressions, contributing to emotional cue extraction for multimodal classification.

The text datasets were merged into a standardized corpus with three target labels: ”depression,” ”stress,” and ”non-mental health.” All datasets were anonymized and publicly available, adhering to ethical guidelines for research involving human data.

Table 4 summarizes the composition and usage of each dataset, including total samples, class distributions, and train/validation/test splits. Stratified sampling was applied to preserve class balance across all subsets. Label values were standardized to binary targets, with 1 indicating the presence of a mental health condition (stress or depression), and 0 representing neutral content. The UCF Selfie Dataset[17] was used solely for validating the image-to-text conversion module and was not used during model training or classification.

Due to limited metadata in the public datasets, demographic attributes such as age, gender, or ethnicity were not consistently available, and subgroup performance analyses were not feasible. This limitation and its implications are addressed in the Bias Mitigation subsection.

Model Selection and Fine-Tuning Process

Seven pre-trained Large Language Models (LLMs) were evaluated for mental health condition detection using zero-shot prompts on multimodal social media data. The models included LLaMA-2-70B-Chat, Mixtral-8x7B-Instruct-v0.1, FLAN-T5-XL, OASST-SFT-1-Pythia-12B, Dolly-v2-12B, Vicuna-13B, and GPT-J-6B (Table 1). These models were selected based on performance in prior NLP tasks and open-access fine-tuning availability. All model selection, fine-tuning, and evaluation were conducted between January and June 2024, based on model availability and infrastructure constraints at that time.

Fine-tuning was conducted over 4,000 training steps, with continuous monitoring of performance metrics on a validation set. A stratified data split was applied, with 70% for training, 15% for validation, and 15% for testing to ensure class distribution was preserved. The final model, Mental-Mixtral, was selected based on its F1-score performance over a 10% validation subset and implemented using QLoRA.

Hyperparameter Optimization and QLoRA Implementation

Quantized Low-Rank Adaptation (QLoRA) was applied to optimize fine-tuning while reducing memory requirements through 4-bit quantization, enabling efficient model training on a single NVIDIA Tesla T4 GPU with 15 GB of memory. The configuration included bnb_4bit_compute_dtype set to float16, bnb_4bit_quant_type set to nf4, and use_nested_quant enabled. Hyperparameter tuning was conducted using a random search strategy across 20 configurations, with the final set selected based on validation performance. The optimal learning rate was 4e-4, chosen from a tested range of 1e-5 to 1e-3, while the weight decay was set to 0.001. A batch size of 4 per device was used, along with a dropout rate of 0.05. For LoRA-specific parameters, the attention dimension (lora_r) was set to 16 and lora_alpha was also set to 16. To further improve memory efficiency, the paged_adamw_8bit optimizer was used in conjunction with a linear learning rate scheduler. The model was trained for up to 4,000 steps, with early stopping employed if no improvement in validation loss was observed. Identical configurations were used across three runs to assess reproducibility (Appendix D).

Evaluation Metrics and Statistical Analysis

Model performance was assessed using standard classification metrics: accuracy, precision, recall, and F1-score. To ensure statistical reliability, 95% confidence intervals (CIs) and Cohen’s d effect sizes were calculated for key metrics. Confusion matrices were generated to visualize classification outcomes (Table 2B, Appendix D).

To validate observed improvements, paired t-tests were conducted comparing F1-scores across three runs of Mental-Mixtral and RoBERTa (text-only baseline).

Multimodal AI Pipeline

The model creation pipeline consisted of five stages: data collection, data preprocessing, image-to-text conversion, model fine-tuning, and classification, ultimately producing the Mental-Mixtral model (Figure 6A). The AI inference pipeline processes multimodal inputs in three sequential phases (Figure 6B). In Phase 1, selfie images were converted into textual descriptions (captions) using LLAVA. In Phase 2, GPT-based models analyzed the combined text from original posts and image-derived descriptions (captions), employing attention mechanisms to highlight key phrases and emotional indicators. In Phase 3, the Mixtral-8x7B-Instruct model classified mental health conditions based on predefined psychological

Table 3: Dataset Composition and Label Distribution

Dataset	Total Samples	Labels (Mental Health = 1, Non = 0)	Class Distribution (%)	Usage	Notes
Reddit Train	10,000	7,804 (1), 2,196 (0)	78.04% / 21.96%	Model training	Combined Reddit stress and depression datasets
Reddit Validation	8,272	4,295 (1), 3,977 (0)	51.93% / 48.07%	Model validation	Balanced validation set from separate subreddit sampling
Multimodal Combined	10,941	6,404 (1), 4,537 (0)	58.52% / 41.48%	Multimodal evaluation	Used to assess impact of LLAVA-generated image descriptions
UCF Selfie Dataset	46,836	N/A (emotion-labeled only)	12-class emotion labels	Image-to-text validation	Not split; used for visual prompt generation with LLAVA

Table 4: Dataset Composition and Label Distribution

Dataset	**Total Samples**	**Labels (Mental Health = 1, Non = 0)**	**Class (%)**	**Usage**	**Notes**
Reddit Train	10,000	7,804 (1), 2,196 (0)	78.04 / 21.96	Model training	Combined Reddit stress and depression datasets
Reddit Validation	8,272	4,295 (1), 3,977 (0)	51.93 / 48.07	Model validation	Balanced validation set from separate subreddit sampling
Multimodal Combined	10,941	6,404 (1), 4,537 (0)	58.52 / 41.48	Multimodal evaluation	Used to assess impact of LLAVA-generated image descriptions
UCF Selfie Dataset	46,836	N/A (emotion-labeled only)	12-class emotion labels	Image-to-text validation	Not split; used for visual prompt generation with LLAVA

indicators. Prompts were structured to guide the model in incorporating both linguistic and visual indicators.

Bias Mitigation Strategies

Potential biases were identified and addressed at each stage of the pipeline. In Phase 1, diverse facial expression datasets representing various ethnicities and cultures were included to mitigate cultural bias in image-to-text conversion. In Phase 2, culturally neutral prompts and diverse linguistic patterns were incorporated to prevent misinterpretation of sarcasm, slang, or colloquial language. Phase 3 involved clinical input to inform classification criteria. Due to missing demographic metadata, subgroup fairness analysis was not performed. Stratified sampling and manual review helped reduce skew. Future work will incorporate demographically labeled data and bias-aware fine-tuning.

Deployment and Application Prototype

The MindWay application prototype (Figure 5) (Figure 6B) was developed to integrate the trained Mental-Mixtral model for real-time adolescent mental health monitoring. The app processes both social media posts and selfies, providing mental health assessments while maintaining data privacy. Anonymization protocols were applied, and opt-out features were incorporated to allow users to manage data usage. Usability testing is planned.

Prompt Design for Mental Health Prediction

Zero-shot prompts were used to instruct the LLMs in classifying mental health conditions from combined text and image inputs [32,33]. These prompts were designed to elicit responses that assess emotional content without prior examples, ensuring generalizability across various data sources and linguistic expressions.

Ethical Considerations

Ethical principles were applied throughout the study. Diverse datasets representing multiple ethnicities, genders, and socioeconomic backgrounds were used to prevent demographic biases. Data privacy and security were maintained by anonymizing all datasets and adhering to General Data Protection Regulation (GDPR) standards. Regular bias assessments were conducted to detect and mitigate any unintended biases in model predictions. Future iterations will integrate federated learning and differential privacy techniques to ensure safety and equity.

ACKNOWLEDGMENTS

We acknowledge Swaminathan Arunachalam and Bhavanthi Thirumanam Raghukumar (Nikila's parents) for their support in getting the tools access, setting up the environment for the experiments, and reviewing the research paper. In addition, we would like to thank Dr. Chris Ng for reviewing and commenting on this research paper.

References

1 J. Śniadach, S. Szymkowiak, P. Osip and N. Waszkiewicz, *Life*, **11**, 1188.
2 *World Health Organization*, www.who.int/news-room/fact-sheets/detail/adolescent-mental-health, Accessed: Jan 30, 2025].
3 M. H. America, Accessed: Jan 30, 2025] .

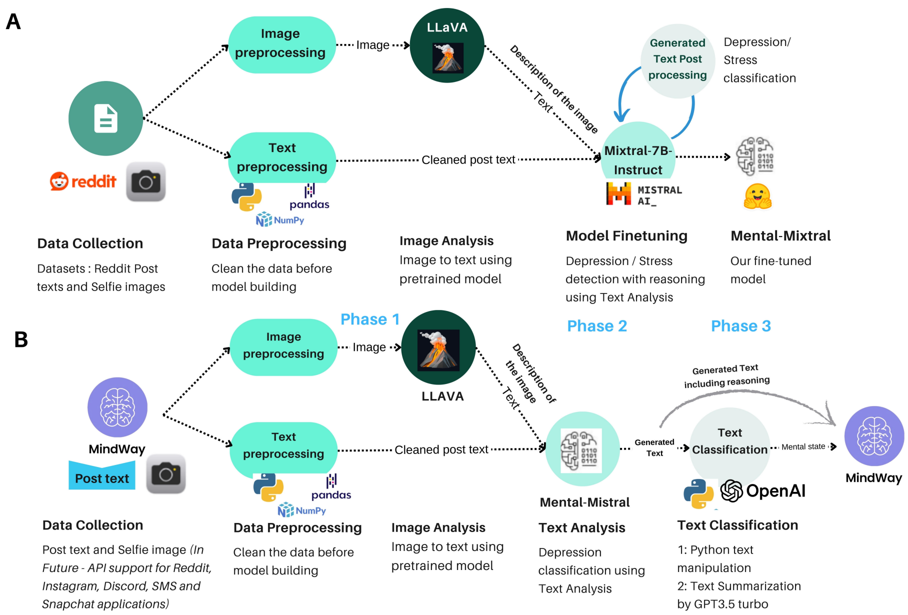

Fig. 6: **AI Pipeline for Mental Health Detection in the MindWay System.** (A) Model creation pipeline showing the progression from data collection and preprocessing to image-to-text conversion, and model fine-tuning, to create the Mental-Mixtral model. (B) Inference pipeline, illustrating three sequential phases: Phase 1 – Conversion of selfie images into descriptive emotional captions using LLAVA; Phase 2 – Fusion and analysis of image-derived captions with text posts using GPT-based models to extract linguistic and emotional features; Phase 3 – Classification of mental health status using the fine-tuned Mixtral-8x7B-Instruct model. This multimodal flow enables real-time prediction of depression and stress within the MindWay application.

4 C. Vidal, T. Lhaksampa, L. Miller and R. Platt, *Int Rev Psychiatry*, **32**, 235–253.

5 K. Smith, D. Ackerman and E. Flanagan, *Compr Psychiatry*, **54**, 1–6.

6 L. Capozzi, *How a lack of rural mental health professionals affects youth in 200 words*, `https://policylab.chop.edu/blog/how-lack-rural-mental-health-professionals-affects-youth-200-words`, Accessed: Jan 30, 2025].

7 *Centers for Disease Control and Prevention (CDC). Child mental health: Rural policy brief*, `https://www.cdc.gov/rural-health/php/policy-briefs/child-mental-health-policy-brief.html`, Accessed: Jan 30, 2025].

8 Increditools, *Teenage use of social media statistics in 2024*, `https://www.increditools.com/teenage-use-of-social-media-statistics/`, Accessed: Jan 30, 2025].

9 M. Anderson and J. Jiang, *Teens and their experiences on social media*, `https://www.pewresearch.org/internet/2018/11/28/teens-and-their-experiences-on-social-media/`, Accessed: Jan 30, 2025].

10 S.A.S., *Natural language processing (NLP): What it is and why it matters*, `https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html`, Accessed: Jan 30, 2025].

11 T. Richter, B. Fishbain, G. Richter-Levin and H. Okon-Singer, *J Pers Med*, **11**, 957.

12 A. Nazir and Z. Wang, *Meta-Radiology*, **1**, 100022.

13 S. Coghlan, J. Reilly and J. Halamka, *Digit Health*, **9**, 1–12.

14 D. Liu, X. Feng, F. Ahmed, M. Shahid and J. Guo, *JMIR Ment Health*, **9, e27244**, year.

15 Z. Cheng, Z.-Q. Cheng, J.-Y. He, J. Sun, K. Wang, Y. Lin, Z. Lian, X. Peng and A. Hauptmann, *Emotion-LLaMA: Multimodal emotion recognition and reasoning with instruction tuning*, arXiv preprint. arXiv:2406.11161.

16 Q. Sun, Y. Cui, X. Zhang, F. Zhang, Q. Yu, Z. Luo, Y. Wang, Y. Rao, J. Liu, T. Huang and X. Wang, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR.

17 *UCF Center for Research in Computer Vision. Selfie dataset*, `https://www.crcv.ucf.edu/data/Selfie/`, Accessed: Jan 30, 2025] .

18 S. Ji, X. Li and Z. Huang, *Neural Comput Appl*, **34**, 10309–10319.

19 D. Owen, E. Gilbert and J. Silva, *JMIR AI*, **2, e41205**, year.

20 A. Aich, A. Vyas and K. Balasubramanian, *Findings EMNLP*, 2871–2887.

21 J. Wei, M. Bosma, V. Zhao, K. Guu, A. Yu, B. Lester, N. Du, A. Dai and Q. Le, *Finetuned language models are zero-shot learners*, arXiv preprint. arXiv:2109.01652.

22 T. Baltrušaitis, C. Ahuja and L.-P. Morency, *IEEE Trans Pattern Anal Mach Intell*, **41**, 423–443.

23 S. Chancellor, M. Birnbaum, E. Caine, V. Silenzio and M. Choudhury, *Proc ACM Conf Fairness Accountability Transparency*, **79–88**, year.

24 Y. Zhao, P. Yin, Y. Li, X. He, J. Du, C. Tao, Y. Guo, M. Prosperi, P. Veltri, X. Yang, Y. Wu and J. Bian, *AMIA Annual Symposium Proceedings*, 1264–1273.

25 A. Caliskan, J. Bryson and A. Narayanan, *Science*, **356**, 183–186,.

26 A. Narayanan and V. Shmatikov, Proc IEEE Symposium on Security and Privacy, p. 111–125.

27 S. Caton and C. Haas, *ACM Comput Surv*, **56**, 1–41.

28 F. Doshi-Velez and B. Kim, *Towards a rigorous science of interpretable machine learning*, arXiv. 1702.08608.
29 E. Turcan and K. McKeown, Proc. Tenth Int. Workshop on Health Text Mining and Information Analysis (LOUHI, p. 97–107.
30 T. Liu, D. Jain, S. Rapole, B. Curtis, J. Eichstaedt, L. Ungar and S. Guntuku, *ACM Web Science Conference*, **15**, 174–183.
31 A. Haque, V. Reddi and T. Giallanza, *Deep learning for suicide and depression identification with unsupervised label correction*, arXiv preprint. arXiv:2102.09427.
32 T. Kojima, S. Sagawa, A. Lu, Y. Li and P. Liang, *Large language models are zero-shot reasoners*, arXiv preprint. arXiv:2205.11916.
33 M. Bosma, J. Wei, V. Zhao, K. Guu, A. Yu, B. Lester, N. Du, A. Dai and Q. Le, *Finetuned language models are zero-shot learners*, arXiv preprint. arXiv:2109.01652.

Appendix A: Overview of MindWay App

As a practical demonstration of our research's innovative multimodal mental health detection model, we developed a prototype application named **MindWay**. This app serves as a proof of concept for using AI in mental health monitoring, leveraging both textual inputs and selfie images to assess the mental health states of adolescents.

Core Functionality

MindWay allows users to manually input text or upload selfies directly through the app interface. The submitted content is processed through the AI pipeline developed in our research, which includes image-to-text conversion and textual analysis using fine-tuned Large Language Models (LLMs). The system assesses whether the content contains indicators of mental health issues such as depression or stress. Results are returned to the user in a non-intrusive and user-friendly manner.

Technical Architecture

The frontend of MindWay was developed using Bubble, a no-code platform, with Figma used for UI design. Screens include user login, data submission (text and images), and results display.

The backend was built using FastAPI to handle user requests and communicate with AI models, hosted on Render. Version control was managed via GitHub. The AI pipeline integrates the LLAVA model for image-to-text conversion and the fine-tuned Mixtral-8x7B-Instruct-v0.1 model for mental health classification. These models are accessed via APIs, with Replicate hosting Mixtral and Hugging Face hosting LLAVA. LangChain manages prompts and model interactions.

Data Flow and Processing

Users authenticate via Bubble's system and can submit text or selfies. Text is tokenized and normalized, while images are resized and passed to LLAVA for conversion to text. Combined prompts are sent to Mixtral via FastAPI, producing 'Yes' or 'No' mental health assessments with brief explanations.

Privacy and Security Measures

Data anonymization removes personal identifiers. Images are not stored post-analysis. All transmissions are HTTPS-encrypted. Users can delete their data at any time. The app complies with GDPR and ethical guidelines.

Future Enhancements

Plans include integrating with social media platforms (Instagram, Snapchat, Discord) via APIs for real-time mental health monitoring, adding anxiety detection, providing resources, and implementing interactive features like mood tracking and personalized feedback.

Appendix B: Effective Prompt Design

Prompt design guided LLMs to classify mental health states accurately from multimodal inputs.

Prompt Template Structure

The structured prompt included:

1. **TextData**: Input textual data or image-to-text conversions.
2. **PromptQ**: Guiding question for mental health prediction.
3. **PromptMS**: Specifies target prediction (e.g., depression or stress).
4. **OutputConstraint**: Ensures output is 'yes' or 'no' with a brief reasoning explanation.

Example Prompt

An example of the structured prompt used in this study is as follows:

TextData: "Here's a social media post: 'Feeling really down lately...' " Posted Image/Selfie: "The person appears with slumped posture..." PromptQ: "Are there discernible signs of depression?" PromptMS: "Depression detection." OutputConstraint: "Provide 'yes' or 'no' and a brief explanation."

Rationale

Clear, structured prompts reduce ambiguity, separate modalities, and ensure consistent output for interpretability. Zero-shot learning was possible with Mixtral-8x7B-Instruct-v0.1 and GPT-3.5 Turbo.

Appendix C: Code Implementation

Repositories and Commit Hashes

MindWay Application

Frontend: Bubble with Figma designs. Backend: FastAPI handling text/image processing, authentication, and session management. AI integration: Mixtral (Replicate) and LLAVA (Hugging Face), managed via LangChain.

Model Building

1. **Image2Text using LLAVA**: Converts selfies to text.
2. **Benchmark LLM Models**: Evaluate multiple LLMs on mental health detection.
3. **Existing Model via Inference API (Mental-RoBERTa)**: Text-based classification.
4. **Fine-Tuning with QLoRA**: Fine-tunes Mixtral-8x7B-Instruct.
5. **Text Classification with GPT-3.5 Turbo**: Detects depression and stress in social media content.

Reproducibility and Environment

Python 3.10, NVIDIA Tesla T4 GPU, CUDA 11.8, packages listed in `requirements.txt`. API keys securely stored.

Appendix D: Statistical Validation and Hyperparameter Optimization

Statistical Validation

Two-tailed paired t-test comparing Mental-Mixtral vs. RoBERTa:

- Mental-Mixtral F1-scores: 92.15%, 92.60%, 91.80%
- RoBERTa: 73.13%, 78.00%, 77.50%
- Mean difference: 14.74
- Standard deviation: 1.41
- Degrees of freedom: 2
- $t(2) = 10.47$, $p = 0.009$
- Cohen's $d = 6.04$

Table 5: Repositories and Commit Hashes for Reproducibility

Component	Repository Name	GitHub URL	Commit Hash / DOI
MindWay App (Backend)	MentalDisorderDetection-BE	GitHub Link	0ca87aa / DOI
Model Building & Benchmarking	DepressionDetection-Model	GitHub Link	e7691f6 / DOI

Table 6: Top 5 Hyperparameter Configurations Ranked by Validation F1-Score

Rank	Learning Rate	Batch Size	Dropout	Weight Decay	LoRA r	LoRA	F1-Score (%)
1	4e-4	4	0.05	0.001	16	16	92.1
2	2e-4	4	0.03	0.001	8	16	91.0
3	6e-4	8	0.05	0.0005	16	32	90.2
4	1e-3	4	0.10	0.0001	8	8	88.9
5	8e-5	8	0.02	0.001	4	16	87.2

Hyperparameter Search and Optimization

Two-phase search: random search followed by grid search, validated on stratified 10% subset. Final configuration: learning rate = 4e-4, batch size = 4, dropout = 0.05.

All models trained for up to 4,000 steps with linear learning rate scheduler, paged_adamw_8bit optimizer, early stopping on plateau. LoRA modules applied with consistent dropout = 0.05.