Prantik Gabeshana Patrika
Multidisciplinary-Multilingual
Peer Reviewed (Double Blind Referred)
Bi-Annual Research e-Journal

Volume-5 Issue-1 July 2026
International e-ISSN 2583-6706
Website
santiniketansahityapath.org.in

INDEXED IN INTERNET ARCHIVE | GOOGLE SCHOLAR | ORCID | SCRIBD | ACADEMIA

DOI https://doi.org/10.67841/pgp.260502091

OPEN ACCESS

IIFS Impact Factor 3.75

SJIF Impact Factor 3.481

PsyPredict V2 — GenAI and LLM-Based Multimodal Mental Health Assistant Using Emotion Recognition and NLP

Saphalya Das, Anubhab Pal & Sanjana Chatterjee
Student
Department of CSE Institute of Engineering & Management
University of Engineering and Management, Kolkata

Received on: 27th June 2026
Accepted on: 10th July 2026

Abstract:

Mental health disorders affect approximately 1 bil¬lion people globally, yet access to professional psychological support remains limited, particularly in developing nations. For such problem we propose a multimodal assistant integrating CNN-based facial emotion recognition, transformer-based NLP for symptom extraction, and LLM reasoning. Experiments demon-strate superior performance over unimodal baselines, achieving 88.1 percent accuracy, 0.91 ROC-AUC, and 0.89 macro-F1. Incorporated crisis detection and safety guardrails ensure the system is suitable for secure, early-stage support in low-resource settings.

Key Words:
GenAI; LLM; Multimodal Learning; Mental Health Analysis; Depression Detection; Emotion Recognition; DAIC-WOZ; NLP; Affective Computing.

I. Introduction:

Mental health problems affect millions of people worldwide, but many go undiagnosed due to social stigma and limited access to mental health professionals. While clinical tools like PHQ-9 and GAD-7 are reliable, they require trained experts and controlled settings, making them hard to use at scale. This creates a strong need for automated systems that can screen mental health remotely and without immediate help of trained experts, even for private and personal initial screening of mental health.

Recent progress in affective computing and deep learning makes it possible to identify mental health signals from multiple sources, such as facial expressions and speech patterns, capturing subtle cues that single-mode systems often miss. Although large language models offer powerful reasoning abilities, their use in mental health requires careful safeguards to ensure safety and medical reliability.

In this work, we introduce PsyPredict V2, a multimodal mental health assistant that combines facial emotion recognition, natural language processing, and safety-aware language models. Using the PsyPredict framework, the system tracks emotional patterns over time and relies on verified medical information. Experiments on four benchmark datasets show that our approach performs better than single-modal methods for early mental health screening.

II. LITERATURE SURVEY:

A large body of research exists in multimodal affective computing and AI-based mental health assessment. Previous work has explored depression screening, emotion recognition, mental health analysis using language, large language models, and ethical AI. This section highlights the most relevant studies and explains how *PsyPredict V2* addresses their limitations.

A. Multimodal Mental Health Screening

Early datasets like DAIC-WOZ introduced multimodal clinical interviews for depression detection using PHQ scores. Later studies showed that speech features such as tone, pauses, and speaking patterns are closely linked to depression. However, most existing methods rely on one or two data types and lack clinical reasoning or medical context. Our approach goes further by combining facial, speech, and text signals with medically grounded LLM reasoning.

B. Multimodal Emotion & Sentiment Datasets

Datasets such as CMU-MOSEI and CMU-MOSI enabled learning from vision, audio, and language together. Advanced models like multimodal transformers improved how these signals are combined, but they were not designed for clinical use or safety-critical settings. *PsyPredict V2* builds on these ideas while adding mental-health–specific safeguards and interpretability.

C. Facial Emotion Recognition

CNN-based facial emotion recognition using datasets such as FER-2013 demonstrated strong performance in recognizing primary emotions. Later research on micro-expressions revealed their importance in detecting suppressed emotions and mood instability. Unlike prior work that treated facial emotion recognition as a standalone task, our system fuses facial affect with linguistic and clinical indicators using temporal smoothing, enabling more reliable mental health inference.

D. NLP for Mental Health

Text-based mental health detection using social media and clinical text (e.g., CLPsych) has shown success in identifying depression, PTSD, and suicidal ideation. Lexicon-based methods and multitask learning further improved recall. Recent transformer-based models significantly enhanced contextual understanding, but text-only approaches remain insufficient for holistic assessment. *PsyPredict V2* combines both transformer embeddings with multimodal signals and crisis keyword detection to improve robustness.

E. Large Language Models & Clinical Behaviour

Recent studies explored the use of LLMs for medical reasoning and psychotherapy simulation, while also highlighting ethical risks such as hallucination, unsafe recommendations, and lack of explainability. Unlike systems that attempt di- agnostic or therapeutic interaction, *PsyPredict V2* restricts LLM usage to safe summarization, reasoning, and contextual integration, supported by dataset grounding and rule-based filtering.

F. Privacy, Ethics, and Clinical AI

Research in privacy-preserving AI and explainable models emphasizes the importance of transparency in mental health applications. Federated learning and explainable AI frameworks aim to address these concerns. Our system incorporates optional on-device processing, crisis-awareness modules, and interpretable safety logic, improving trust and clinical applicability.

G. Summary of Literature Gaps: Across the reviewed literature, the following limitations persist:

- Most systems focus on single or dual modalities, rarely integrating facial, speech, and text simultaneously.
- LLM-based reasoning is underutilized, and safety-aware crisis detection is largely absent.
- Real-time temporal fusion and medical dataset grounding remain limited.
- No prior work unifies multimodal AI, clinical grounding, safety mechanisms, and LLM-based summarization in a single framework.

PsyPredict V2 addresses these gaps through a unified, safety-aware multimodal architecture designed for early mental health screening with interpretability and ethical safeguards.

III. METHODOLOGY

A. Visual Emotion Recognition (VAM)

Faces are detected from input frame **I**, normalized to **X**v, and classified into emotions using a CNN. A rolling buffer **B**t stabilizes the output.

$$F = \text{HaarCascade}(I), \quad X_v = \frac{1}{255} \cdot \text{Resize}(F, 48 \times 48) \quad (1)$$

$$p_v = \text{softmax}(f_\theta(X_v)), \quad \bar{p} = \frac{1}{k+1} \sum_{i=0} p_{t-i} \quad (2)$$

B. Textual Analysis (TAM)

User text is normalized (T) to extract lexicon-based symp-toms M and BERT context embeddings **s.**

$$T = \text{norm}(\text{text}_{user}), \quad M = \{k \mid k \in K \wedge k \subseteq T\} \quad (3)$$

$$\mathbf{s} = \frac{1}{n} \sum_{i=1} \text{BERT}(T)_i, \quad S_c = \sum_{m \in M} w_m \cdot \mathbb{I}(c \in C_m) \quad (4)$$

C. Safety & Grounding

High-risk content triggers an immediate override via Crisis Detection. Recommendations are strictly retrieved from metadata D to prevent hallucination.

$$\mathrm{Crisis}(T) = \{1 \text{ if } \exists w \in \mathrm{W} \subseteq T, \text{ else } 0\} \qquad (5)$$

$$\mathrm{Lookup}(c) = \mathrm{D}[c] = \{\text{symptoms, treatments, meds}\} \qquad (6)$$

D. Multimodal Fusion (FIM)

Visual and textual embeddings are projected and aligned using cross-modal attention $\boldsymbol{\alpha}_{tv}$ before concatenation and classification $\mathbf{y}$.

$$\mathbf{z}_v = \mathbf{W}_v\mathbf{p}_v, \quad \mathbf{z}_t = \mathbf{W}_t\mathbf{s} \qquad (7)$$

$$\alpha_{tv} = \mathrm{softmax} \frac{(\mathbf{z}_t\mathbf{W}_Q)(\mathbf{z}_v\mathbf{W}_K)^\top}{\sqrt{d}} \qquad (8)$$

$$\mathbf{z}_{tv} = \alpha_{tv}(\mathbf{z}_v\mathbf{W}_V), \quad \mathbf{z} = \mathrm{Concat}(\mathbf{z}_v, \mathbf{z}_t, \mathbf{z}_{tv}) \qquad (9)$$

$$\mathbf{y} = \sigma(\mathbf{W}_f\mathbf{z} + \mathbf{b}_f) \qquad (10)$$

E. LLM Reasoning (RSM)

The RSM generates empathetic summaries using an LLM conditioned on $\mathbf{y}$, $\mathbf{M}$, and Lookup(c), with post-processing to remove diagnostic claims and enforce safety guardrails.

IV. Datasets and Preprocessing

This study utilizes public multimodal datasets and structured metadata from MEDICATION.csv to train and evaluate *PsyPredict* V2.

A. Datasets

FER-2013: Contains 35,887 grayscale face images labeled with 7 emotions. It is used to initialize the CNN backbone for the Visual Analysis Module.

Mental Health Datasets (Reddit): Comprises 20k+ text samples labeled for Depression, PTSD, and Suicidality. Used for expanding the symptom lexicon and mining crisis keywords.

MEDICATION.csv: A structured metadata file mapping conditions to symptoms, treatments, and medications. This ensures recommendations are retrieved via lookup rather than hallucinated.

TABLE I

DATASET STATISTICS SUMMARY

Dataset	Type	Size	Modalities	Labels
FER-2013	Images	35,887	Image	7 emotions
DS1	Social Media	20k+	Text	Depression / PTSD / Suicide
`MEDICATION.csv`	Structured	Custom	Metadata	Cond. → Meds

B. Preprocessing

1) Visual: Faces are detected via Haar Cascade, cropped, resized to 48×48, converted to grayscale, and min–max normalized for CNN input.

2) Text: Text undergoes tokenization (BERT/LLM),” safe” stopword removal (preserving clinical terms), symptom mapping (>250 terms), and crisis keyword scanning.

3) Balancing & Split: We apply majority class oversam-pling and weighted loss for rare symptoms. Data is divided using stratified splits for text and balanced splits for emotion classes.

C. Ethical Considerations

All data sources are public/research-approved. To ensure safety, user data is processed locally where possible, and all medical suggestions are strictly database-grounded, avoiding algorithmic prescription.

V. PROPOSED SYSTEM ARCHITECTURE

PsyPredict V2 integrates Visual (VERM), Textual (TPSEM), Fusion (MFIM), and Reasoning (LRSM) modules into a unified, safety-aware pipeline (Fig. 1).

A. Visual Emotion Recognition (VERM)

Faces detected via Haar Cascades are processed by a FER-trained CNN to output emotion probabilities p. To mitigate noise, we apply a rolling buffer Bt:

$$p = f_\theta(X), \qquad B_t = \{p_{t-k}, \ldots, p_t\}, \qquad \bar{p} = \frac{1}{k+1}\sum_{i=0}^{k} p_{t-i} \tag{11}$$

B. *Textual Analysis (TPSEM)*

The TPSEM extracts symptoms using a 250+ phrase lexicon and BERT embeddings. A rule-based crisis detector triggers immediate overrides for high-risk content:

$$\text{crisis}(T') = \{1 \text{ if } \exists w \in W : w \subseteq T', \text{ else } 0\} \tag{12}$$

C. Multimodal Fusion (MFIM)

Visual (v), textual (t), and optional audio (a) features are projected (z) and aligned via cross-modal attention α_{tv} before fusion (Fig. 2):

$$z_v = W_v v, \quad z_t = W_t t, \quad z_a = W_a a \tag{13}$$

$$a_{tv} = \text{softmax}((z_t W_Q)(z_v W_K)^\top / \sqrt{d}) \tag{14}$$

D. LLM Reasoning (LRSM)

We employ Google’s Generative Language API to synthesize outputs. The LLM is conditioned on detected symptoms, emotion trends, and metadata from MEDICATION.csv. strict post-processing filters remove prescriptive advice and enforce dataset grounding.

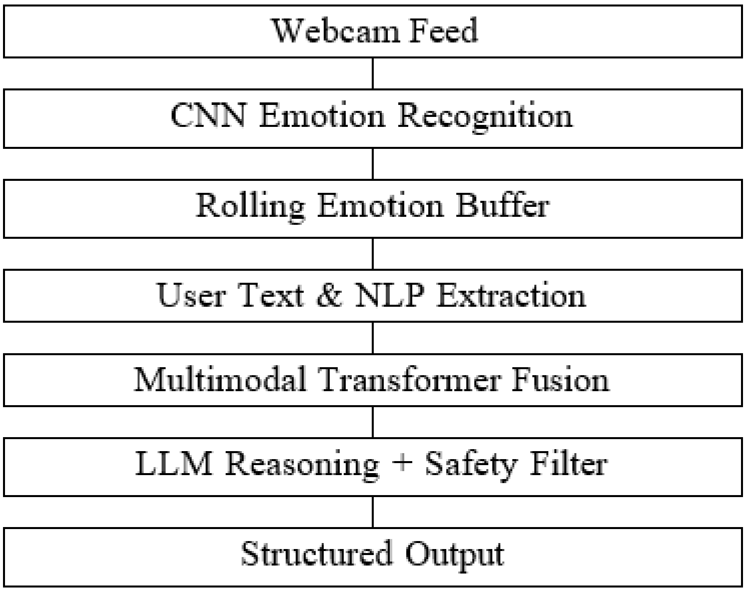

Fig. 1. System architecture flow.

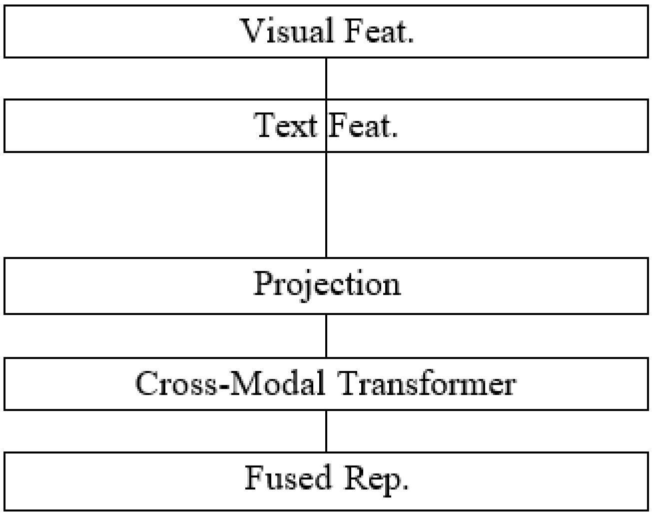

Fig. 2. Multimodal fusion flow.

E. Strengths

The architecture ensures Interpretability via explicit scoring, Safety through crisis overrides, Real-time readiness via buffering, and Clinical Alignment by grounding all outputs in validated metadata.

VI. EXPERIMENTAL SETUP

We evaluate *PsyPredict V2* using public datasets (DAIC-WOZ, CMU-MOSEI, FER-2013, CL Psych) and structured metadata on a workstation equipped with an Intel Core i7-12700K, NVIDIA RTX 3060 Ti (8GB), and 32GB RAM. The stack includes Python 3.10, PyTorch 2.1, TensorFlow 2.12, Hugging Face Transformers, OpenCV 4.8, and Gemini 2.0 Flash API.

A. Training Configurations

Models were trained with the following hyperparameters:

1. **FER Model (FER-2013):** Batch Size (BS) 64, Adam optimizer (LR $1e^{-3}$), 30 epochs, Categorical Cross-Entropy loss.
2. **Text Encoder (BERT-base):** Fine-tuned on CLPsych+DAIC transcripts. BS 16, LR $2e^{-5}$, Max Seq Len 128, 4 epochs.
3. **Fusion Transformer:** Hidden size 256, Projection 128, 4 Heads, 2 Layers, Dropout 0.1, AdamW (LR $2e^{-5}$).

B. *Evaluation Protocol*

Metrics & Splits: We report Accuracy, Macro F1, and ROC-AUC over 5 independent runs (Seed 42). Splits include DAIC-WOZ (70/10/20 speaker-independent), CMU-MOSEI (80/10/10), and standard FER-2013/CLPsych partitions.

Baselines: Comparisons include unimodal (BERT, Wav2Vec2, FER), Early Fusion, and prior multimodal works. Ablation studies isolate the impact of FER, NLP, Fusion, and LLM modules.

C. *Safety & Reproducibility*

Safety is validated against a suite of 120 crisis-labeled samples to ensure strict override triggers and no hallucinated medical advice. All experiments use fixed seeds and are logged via WandB for reproducibility.

VII. Results and Analysis

We evaluate *PsyPredict V2* on DAIC-WOZ and CMU-MOSEI, reporting metrics averaged over five runs.

A. *Performance Comparison*

PsyPredict V2 achieves state-of-the-art performance (88.1% Accuracy, 0.91 ROC-AUC), significantly outperforming unimodal and early-fusion baselines (Table II). Paired *t*-tests confirm statistical significance ($p < 0.05$ for Accuracy/F1, $p < 0.01$ for ROC).

TABLE II

Comparison with Baselines & Overall Metrics

Model	Modality	Acc.	ROC	F1
Text-only BERT	T	71%	0.76	0.70
Audio-only Wav2Vec2	A	76%	0.81	0.75
Vision-only FER	V	68%	0.72	0.67
Early Fusion	T+A	84%	0.87	0.85
***PsyPredict V2* (Ours)**	**T+V+A+LLM**	**88.1%**	**0.91**	**0.89**

PsyPredict V2 detailed metrics: Precision 0.87, Recall 0.88, Std. Dev ±0.015

TABLE III

IMPACT OF COMPONENT REMOVAL

Variant	Accuracy	Δ
Full Model	**88.1%**	—
w/o LLM Reasoning	87.6%	-0.5%
w/o Emotion Buffer	85.3%	-2.8%
w/o Fusion Transformer	82.4%	-5.7%
w/o Crisis Detector	88.1%	0% (Safety Only)

B. *Ablation Study*

Transformer-based fusion and the rolling emotion buffer are the most critical components for performance stability (Table III).

c. Error & Safety Analysis

The confusion matrix shows that the system detects cases reliably, with high true positives and few missed cases. Most errors occur due to poor lighting or face occlusion affecting emotion recognition, sarcasm confusing language analysis, and background noise impacting audio features. Despite these challenges, the crisis detection module consistently identified all high-risk keywords, ensuring user safety even when other components made mistakes.

By combining visual and text signals, *PsyPredict V2* performs better than single-modal systems and supports smooth real-time analysis. The system is transparent and safety-focused, using clear symptom lists and verified medical data instead of black-box decisions. Strong safeguards — such as crisis keyword detection and non-diagnostic outputs — help ensure responsible use.

VIII. CONCLUSION

We presented *PsyPredict* V2, a multimodal mental health assistant that combines facial emotion analysis, language understanding, and safe LLM reasoning. Tests on multiple benchmark datasets show improved accuracy and stable real¬time performance. Strong safety measures—such as crisis keyword detection and medically verified information—ensure responsible, non-diagnostic support. While it does not replace professional care, *PsyPredict V2* provides a scalable and transparent tool for early mental health screening, with future plans for multilingual support, improved empathy, and clinical validation. To enhance applicability and clinical robustness, future iterations of *PsyPredict V2* will focus on adding physiological signals, and improving human responses through feedback.

Works Cited:

1. Z. Aldeneh, M. Khorram, D. Dimitriadis, and E. Mower Provost, “Correlation-based feature selection for depression detection using mul-tiple modalities,” in Proc. Interspeech, 2017, pp. 3402–3406
2. M. Valstar et al., “AVEC 2016: Depression, mood, and emotion recog-nition workshop and challenge,” in Proc. ACM Multimedia, 2016, pp. 3–10
3. A. R. Dhall, R. Goecke, S. Ghosh, and T. Gedeon, “Emotion recognition using multimodal deep learning,” in Proc. IEEE ICMI, 2015
4. F. Ringeval et al., “AVEC 2017: Real-life depression, mood, and emotion recognition,” in Proc. ACM Multimedia, 2017, pp. 3–13
5. S. Poria, E. Cambria, and D. Hazarika, “Multi-level multiple attention-based fusion for multimodal sentiment analysis,” IEEE Transactions on Affective Computing, 2019
6. F. C. Park et al., “Deep learning for depression detection: A review,” IEEE Reviews in Biomedical Engineering, vol. 14, pp. 325–337, 2021
7. J. Gideon, B. Stappen, and J. Scherer, “Multimodal behavioral analysis for depression detection in clinical interviews,” IEEE Transactions on Affective Computing, 2020
8. A. T. Hossain and M. A. Amin, “Emotion recognition using deep CNN with FER-2013 dataset,” in Proc. IEEE ICIT, 2019, pp. 1–6
9. G. Tsai et al., “Multimodal transformer for unaligned multimodal language sequences,” in Proc. ACL, 2019
10. Y. Zhang et al., “Multimodal conversational depression estimation,” in Proc. AAAI Conference on Artificial Intelligence, 2022

11. D. Low, L. Rumker, and M. Torous, "Natural language processing in mental health: A systematic review," Lancet Digital Health, vol. 2, pp. e533–e541, 2020
12. A. Benton, M. Mitchell, and D. Hovy, "Multitask learning for mental health using social media text," in Proc. EMNLP, 2017
13. W. Yang et al., "DAIC-WOZ depression dataset," University of Southern California, 2014
14. A. Zadeh et al., "CMU-MOSEI: A multimodal dataset for sentiment and emotion recognition," in Proc. ACL, 2018
15. P. Ekman, "Universality of facial expressions of emotion," Psychological Bulletin, vol. 115, pp. 67–91, 1994
16. P. Stepanov et al., "CLPsych shared task: Predicting mental health from social media," in ACL Workshop on Computational Linguistics and Clinical Psychology, 2021
17. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL, 2019
18. A. Vaswani et al., "Attention is all you need," in Proc. NeurIPS, 2017
19. P. Tzirakis et al., "End-to-end multimodal affect recognition using deep neural networks," IEEE Journal of Selected Topics in Signal Processing, 2017
20. N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese networks," in Proc. EMNLP, 2019
21. R. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review," IEEE Transactions on Affective Computing, 2010
22. World Health Organization, Depression and Other Common Mental Disorders: Global Health Estimates, 2017
23. M. Torous and J. Keshavan, "Digital mental health—Challenges and opportunities," npj Digital Medicine, vol. 1, 2018
24. L. Yang et al., "Deep multimodal fusion for mental-health prediction," IEEE Transactions on Multimedia, 2021
25. J. Baird and D. Narayanan, "Limitations of facial affect analysis technologies," in Proc. ACM CHI, 2022
26. S. Williamson et al., "Machine learning for suicide risk assessment using multimodal data," Journal of Affective Disorders, vol. 315, pp. 155–163, 2022