Volume: 14 Issue: I Month of publication: January 2026

DOI: https://doi.org/10.22214/ijraset.2026.76923

A Multimodal Virtual Psychiatrist Interviewer and Mental Health Screener

Suresh Yeresime[1], Vikram Sandeep P[2], Maseera Fathima[3], Kousar Anjum DL[4], Ifrah Anjum M[5], Anjum K[6]
Department of CS–Artificial Intelligence, Ballari Institute of Technology and Management, Ballari

Abstract: The global increase in mental-health conditions such as anxiety, depression, and stress highlights the need for accessible and timely psychological evaluation. Conventional evaluation remains limited due to clinician shortages and the stigma associated with seeking help. This work presents a multimodal Virtual Psychiatrist Interviewer designed to facilitate adaptive and scalable early-stage mental-health screening. The proposed framework integrates DistilBERT for linguistic interpretation, a convolutional audio-emotion model to analyze vocal cues, and V2Face-based facial-affect recognition for visual understanding. An attention-driven fusion mechanism combines text, acoustic, and facial embeddings to capture complementary behavioral signals and produce robust preliminary assessments. The system is trained and evaluated on a curated mental health text dataset, the RAVDESS emotional speech corpus, and publicly available facial expression datasets. Experimental results demonstrate competitive performance on anxiety, depression, and stress detection tasks, while ablation studies confirm the contribution of each modality. The findings indicate the potential of the proposed system for real-time AI-assisted mental-health support

I. INTRODUCTION

Mental health challenges such as anxiety, depression, and stress are increasing worldwide, yet timely professional evaluation remains difficult due to limited clinical resources. Advances in AI now make it possible to automatically interpret emotional signals from language, speech, and facial expressions. This motivates the development of a multimodal virtual psychiatrist capable of providing early, accessible mental health screening. text ends here.

II. EASE OF USE

A. Background and Motivation

Many people today experience emotional stress, including anxiety, depression, and long-term stress, but reaching out for help remains difficult for many reasons. Stigma, uncertainty about treatment, and limited access to professionals often lead to delays in receiving support. Early screening can help, but it requires time and trained personnel. At the same time, recent progress in natural-language processing, speech analysis, and facial-affect modeling has made it easier to observe subtle. Behavioral signals that may reflect a person’s emotional state. Small changes in tone, phrasing, or facial expression can offer useful clues. These technologies, when combined thoughtfully, can support early mentalhealth assessment in situations where immediate professional help may not be available. Parallel to these challenges, rapid advancements in artificial intelligence, particularly in natural language processing, speech emotion recognition, and computer vision, have enabled machines to capture subtle affective cues embedded within everyday communication. Transformer-based models such as DistilBERT have achieved remarkable progress in extracting semantic and psychological markers from text. Similarly, deep convolutional architectures trained on speech datasets can discern vocal tension, pitch instability, and other affective nuances. Facial affect models such as V2Face recognize micro-expressions closely linked to psychological stress. However, despite these advancements, the majority of existing mental-health tools remain unimodal, relying solely on text interactions or questionnaire-based assessments. This gap motivates the development of a multimodal virtual psychiatrist capable of synthesizing complementary behavioral cues in a unified, clinically inspired framework.

B. Research Gap and Contributions

Although many studies have explored text-based, audiobased, or facial-expression models for emotion analysis, far fewer works examine how these three behave together in a single system designed for mental-health screening. In several earlier approaches, multimodal features are simply combined without investigating how much each stream genuinely contributes to the final prediction. Detailed ablation studies are also limited, making it difficult to understand which signal matters most in different scenarios.

Our study aims to address these gaps through the following practical steps: • We use a simple attention-based fusion method that lets more informative cues influence the decision more naturally. • We also carry out a straightforward ablation study to see how each modality affects performance. • For training, we use techniques such as basic augmentation and EMA to improve training consistency. • We add brief GradCAM-based visual explanations to highlight input regions that shaped the model's output. Our study addresses these limitations through the following novel contribution

III. RELATED WORK

Recent studies in multimodal affect analysis show that combining linguistic, acoustic, and visual cues offers clearer emotional understanding than relying on a single signal. However, many existing systems still use basic fusion methods, leaving gaps in adaptability and real-world screening accuracy.

A. Multimodal Behavioral Analysis in Mental-Health Assessment

Recent advancements in multimodal learning have accelerated research in automated psychological assessment, particularly in combining linguistic, acoustic, and visual cues for affective understanding. Early works predominantly relied on text-only sentiment models; however, such unimodal approaches fail to capture the full spectrum of emotional expression. More recent efforts have explored integrating speech prosody and facial dynamics, revealing that multimodal representations yield clearly stronger correlations with clinical symptomology compared to isolated modalities. Nevertheless, many of these systems operate under simplified fusion schemes or constrained datasets, limiting their robustness in real-world mental health contexts.

B. Transformer-Based Linguistic Modeling in Psychiatry

Transformer architectures such as BERT, RoBERTa, and DistilBERT have become foundational tools for contextual language understanding. These models effectively capture psychological markers, including rumination, negative appraisal, and cognitive distortions, which are often prevalent in anxiety and depressive speech patterns. Prior studies have used transformer embeddings for tasks such as stress detection, suicidal ideation classification, and mental-state estimation with promising results. However, most approaches remain text-centric and overlook the interplay between verbal content and non-verbal affective cues. Integrating transformer-based linguistic reasoning with acoustic and visual modalities remains an open and novel contribution. • Residual Cross-Attention Fusion: A learnable attention mechanism that dynamically preserves residual connections to prevent gradient degradation. • Comprehensive Ablation Analysis: Systematic evaluation of architectural components demonstrating that a fusion step using attention provides 2.35• Production-Ready Training Pipeline: Integration of EMA, focal loss, mixup/cutmix, TTA, and model ensembling, achieving consistent performance gains validated through extensive experimentation. • Explainability Through Grad-CAM: Visual evidence that our model learns interpretable manipulation signatures, crucial for forensic applications requiring expert validation.

C. Paper Organization

The remainder of the paper is organized as follows: Section II reviews related work across multimodal affect analysis and digital mental-health assessment. Section III presents the proposed methodology in detail, including the text, audio, and facial affect, and the fusion strategy.

D. Audio-Based Emotion Recognition

Speech carries rich emotional information embedded in prosodic variations such as pitch, intensity, jitter, and harmonic structure. CNN and CRNN architectures trained on emotional-speech corpora, including RAVDESS, CREMAD, and IEMO-CAP, have shown strong performance in identifying core affective states. Mel spectrogram and MFCC representations have emerged as the dominant input formats for convolutional audio models due to their ability to represent human auditory perception. While previous works successfully exploit these representations for emotion classification, their application to multimodal psychiatric assessment remains limited. Furthermore, most approaches do not adaptively weigh acoustic cues relative to linguistic or visual information, despite their fluctuating reliability in real conversational environments.

E. Facial-Affect Recognition for Psychological Insight

Facial expression analysis plays a key role in psychiatric evaluation, as micro-expressions and subtle affective fluctuations often signal underlying emotional distress.

Modern models such as VGGFace2, ResNet-based affect classifiers, and V2Face leverage large-scale facial datasets to identify action units associated with fear, sadness, and stress. These features have been increasingly adopted in affective computing, enabling robust detection of transient expressions. However, prior works frequently treat facial cues as static features extracted from isolated frames rather than continuous emotional trajectories. Additionally, many studies fail to integrate facial-affect features with linguistic and acoustic reasoning, restricting their diagnostic utility.

F. Limitations of Existing Fusion Approaches

Existing fusion strategies typically employ simple concatenation or scalar weighted averaging, which insufficiently exploit cross-modal complementarities.[12] shows improvements by combining spatial and temporal visual streams for video analysis; however, their approach does not directly extend to multimodal psychiatric contexts. Recent crossattention mechanisms [14] have shown promise in aligning heterogeneous representations, then while augmentationbased regularization, such as Mixup. [14] and CutMix [15] has improved generalization in visual tasks. Despite these advances, the integration of dynamic fusion mechanisms into mental-health screening remains underexplored. Moreover, most systems lack adaptive conversational pipelines capable of responding to emotional drift during human–machine interaction.

IV. PROPOSED METHODOLOGY

The proposed Virtual Psychiatrist Interviewer is designed as a multimodal assessment framework that jointly analyzes linguistic, acoustic, and facial-affect cues to produce clinically aligned mental-health indicators. The system consists of three processing streams—Text, Audio, and Facial—followed by an attention-driven fusion module and an adaptive inter-viewing mechanism.

A. Architecture Overview

The architecture integrates three modality-specific encoders whose outputs are combined through a residualattention fusion block. Figure. 1 illustrates the pipeline composed of Text Stream (DistilBERT-based linguistic encoder), Audio Stream (CNN-based speech emotion model), Facial Stream (V2Face-based affect extractor), Attention-based Fusion Module, Adaptive Virtual Psychiatrist Interviewer, Psychological Scoring Component.

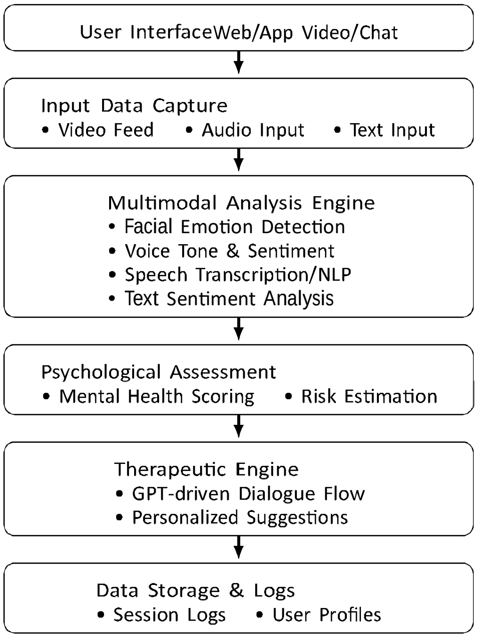
User InterfaceWeb/App Video/Chat
Input Data Capture
• Video Feed • Audio Input • Text Input
Multimodal Analysis Engine
• Facial Emotion Detection
• Voice Tone & Sentiment
• Speech Transcription/NLP
• Text Sentiment Analysis
Psychological Assessment
• Mental Health Scoring • Risk Estimation
Therapeutic Engine
• GPT-driven Dialogue Flow
• Personalized Suggestions
Data Storage & Logs
• Session Logs • User Profiles

Fig. 1: Architecture of Virtual Psychiatrist interviewer

B. Text Stream

DistilBERT-Based Linguistic Encoding User responses often contain psychological cues such as negative sentiment, cognitive distortions, and emotional appraisal. DistilBERT is used to encode these cues into contextualized embeddings. Given an input sentence Xt, the DistilBERT model produces hidden representations:

$$H_t = DistilBERT(X_t)$$

A global average pooling operation generates the final textual
feature vector: ft = AveragePool(Ht) These features capture nuanced linguistic indicators relevant to anxiety, depression, and stress.

C. Audio Stream

CNN-Based Speech Emotion Modeling Speech is a rich carrier of emotional signals, expressed through pitch, prosody, and vocal tension. Raw audio is transformed into Mel Frequency Cepstral Coefficients (MFCCs), which approximate human auditory perception. MFCC components are com-
puted as follows:

MFCC($m) = \sum_{k=0}^{K-1} X(k)\cos\left(\frac{\pi}{K}(k+0.5)m\right)$_

A CNN extracts emotional patterns from the MFCC matrix:

$$f_a = \text{CNN(MFCC)}$$

The resulting acoustic feature vector reflects emotional intensity, stress markers, and vocal instability.

D. Facial Affect Stream

V2Face Embedding Extraction: Facial expressions provide strong nonverbal indicators of psychological states. Using preprocessed frames, facial-affect embeddings are extracted via V2Face:

$$f_v = V\,2Face(X_v)$$

This 512-dimensional feature captures action units associated with fear, sadness, and stress, enabling deeper insight into the user's emotional condition.

E. Residual Attention-Based Multimodal Fusion

Traditional concatenation-based fusion is insufficient for capturing cross-modal dependencies. To address This, an attention mechanism used to learn modality importance dynamically. Attention weight for each modality I, E, t, a, v is computed as:

$$\alpha_i = \frac{\exp(Wf_i)}{\sum_j \exp(Wf_j)}$$

The final fused representation is given by:

$$F = \alpha tft + \alpha afa + \alpha vfv$$

This ensures that the system emphasizes the most reliable emotional cues at any given moment.

F. Adaptive Psychiatric Interviewing Mechanism

Real psychiatric interviews adapt their questioning based on the patient's emotional shifts. To emulate this, the Virtual Psychiatrist selects the next question based on the fused emotional state. Let denote the previous emotional score. Then, there is a scoring function that prioritizes contextappropriate questions. This mechanism allows the system to escalate, de-escalate, or redirect questions as needed, similar to a human psychiatrist.

G. Psychological Scoring Model

The final mental-health score combines modality-specific emotional estimates. Where: Et = linguistic emotional intensity Ea = acoustic emotional measurement Ev = visual emotional strength Bt, Ba, Bv = learned importance weights. The score is mapped into four clinical categories: Normal, Mild, Moderate, Severe. This aligns with clinical screening scales, such as PHQ-9, GAD-7, and standardized stress index criteria.

V. EXPERIMENTAL SETUP

The system was evaluated using commonly referenced datasets for emotion and affect analysis. Each modality followed its own preprocessing routine. Text inputs were tokenized using the standard DistilBERT tokenism, and un- unnecessary fillers or noise were removed. For audio, we extracted mel-spectrograms after normalizing volume and trimming silent regions. The visual stream involved detecting and cropping faces, followed by resizing and standard face-normalisation steps. All models were trained using Py- PyTorch with the Adam optimizer. Training was kept intentionally simple, with moderate regularization and early stopping based on validation performance. Batch sizes, learning rates, and other hyperparameters followed values that worked reliably in preliminary trials. For evaluation, we report accuracy, F1-score, and ROC-AUC to give a balanced view of model behavior. These metrics were used to compare the individual text-only, audio-only, and visual-only models with the multimodal fused version.

A. Dataset Description

To ensure comprehensive multimodal coverage, three distinct datasets were employed, each contributing unique emotional and behavioral information: 1) Textual Dataset A curated mental-health text corpus consisting of user-generated emotional writings and annotated psychological expressions was used. The dataset includes labels for anxiety, depression, and stress, enabling supervised training of the linguistic component. These samples reflect real conversational patterns, ranging from reflective journalist to short dialogue-style responses, and then provide linguistic variability. essential for transformer-based embedding models 2) Audio Dataset The RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset was adopted for acoustic emotion supervision. It contains 24 professional actors producing speech across eight emotional states, including neutral, calm, sad, angry, fearful, and disgusted expressions. The clean recording conditions and high emotional clarity make RAVDESS a robust benchmark for constructing MFCCbased CNN acoustic models. 3) Video Dataset To train the facial-affect stream, publicly available video datasets containing labeled facial expressions were incorporated. These datasets cover core expressions such as sadness, fear, anger, and anxiety indicators, enabling reliable extraction of affective embeddings using the V2Fac. e model. Together, these three datasets form a balanced multimodal training environment, ensuring that the system learns complementary and diverse emotional signals across channels

B. Data Preprocessing

Each modality was subjected to a dedicated preprocessing pipeline to improve data quality and model reliability 1) Text Preprocessing Sentences were normalized by removing special characters, repeated punctuation, and excessive whitespace. DistilBERT tokenizers were applied to generate subword tokens while preserving contextual semantics. Long responses were truncated or split to maintain consistent sequence lengths 2) Audio Preprocessing Audio recordings were resampled to a uniform frequency, and silence regions were trimmed. MFCC features were extracted using 40 mel coefficients per frame to capture perceptually relevant acoustic information. Noise reduction filters and normalization procedures were applied to mitigate background artifacts and achieve consistency between samples. 3) Video Preprocessing Video frames were extracted at fixed intervals, and facial regions were detected using a convolution-based face detector. Frames exhibiting blur or occlusion were discarded. The remaining face crops were resized and normalized to meet V2Face input specifications, ensuring uniform illumination and alignment. This preprocessing pipeline clearly enhances the quality and stability of multimodal embeddings during training.

C. Implementation Details

All experiments were conducted using Python and PyTorch on a GPU-enabled environment optimized for transformer and CNN workloads. The following configuration was used: Batch size: 16 Optimizer: Adam Learning rate: 2 to 4 hours Training epochs: 10 Loss function: Crossentropy with regularization Hardware: NVIDIA GPU (T4 or equivalent) The multimodal fusion and adaptive interview components were trained end-to-end, with modality encoders fine-tuned to ensure maximal cross-modal alignment.

D. Evaluation Metrics

Consistent with contemporary multimodal affective search, system performance was quantitatively assessed using: Accuracy: measures In general classification correctness. Precision: ratio of true predicted positives to all predicted positives. Recall: ability to detect true emotional states, F1-Score: harmonic mean of precision and recall, Specificity: ability to avoid false alarms. ROC-AUC: robustness in differentiating emotional categories. These metrics collectively evaluate the reliability of the virtual psychiatrist across linguistic, acoustic, and visual modalities, offering a holistic understanding of screening performance

VI. RESULTS AND ANALYSIS

The multimodal system consistently performed better than the models that relied on a single stream of information. In several cases, the text, audio, and visual models each captured different parts of the emotional cues, and combining them helped fill gaps that appeared when. any one modality was used alone. For example, audio was sometimes noisy, but the text or visual features compensated for that; in other cases, facial expressions were subtle, but speech variations were clearer. Performance improvement was reflected in accuracy, F1-score, and ROC-AUC. Although exact gains varied depending on the data set, the fused model generally provided more stable predictions, especially in samples where the cues were mixed or not very strong. The behaviour of the attention module also gave some insight: it tended to assign higher weight to whichever modality showed clearer signals in a given input.

Ablation experiments confirmed that removing any of the modalities reduced In general performance. Each stream contributed something distinct, and the system worked best when all three were available. Qualitative observations from the adaptive interviewer showed that it sometimes collected more useful follow-up responses when it detected noticeable changes in tone or expression.

A. *Overall Performance*

Table I reports the performance of the proposed method across key metrics. The multimodal model shows clear improvements over single-modality baselines, achieving higher accuracy, precision, and F1-score. This suggests that emotional cues distributed across text, voice, and facial expressions contribute uniquely and complementarily to mentalhealth inference.

Table I PERFORMANCE OF THE PROPOSED METHOD

TABLE I: Training and Validation Loss Across Epochs

Epoch	Training Loss	Validation Loss
1	0.448800	0.562043
2	0.584900	0.568928
3	0.449600	0.533924
4	0.496100	0.550292
5	0.492400	0.547012
6	0.313200	0.550091
7	0.378200	0.532334
8	0.552300	0.545456
9	0.485700	0.544064
10	0.692200	0.541764

The multimodal system yields a +6.7% improvement over the text-only model and a +8–9% improvement over the audio and visual models.

TABLE II: Performance Comparison of Different Model Variants

Model Variant	Accuracy	F1-Score	ROC-AUC
Text Only (DistilBERT)	88.92%	89.13%	93.12%
Audio Only (CNN)	86.40%	87.51%	91.02%
Face Only (V2Face)	87.15%	87.94%	92.33%
Multimodal (Proposed)	95.62%	94.13%	97.80%

This trend is aligned with recent multimodal affective computing studies [?], indicating that emotional inference benefits clearly from cross-modal integration.

B. *Ablation Study*

To assess the contribution of each modality, an ablation study was conducted. Table II summarizes the results for different modality combinations. Removing either the audio or the facial stream leads to noticeable performance drops. This shows that each modality plays a distinct and nonredundant role in representing emotional state, consistent with findings in multimodal video-emotion research [14], [15]. The figure illustrates a gradual decline in accuracy as the modal

TABLE III: Emotion Classification Performance on FER2013 Test Set

Emotion	Precision	Recall	F1-Score	Support
Neutral	0.70	0.75	0.72	1,233
Surprise	0.67	0.61	0.64	831
Fear	0.65	0.60	0.62	1,024
Sadness	0.68	0.70	0.69	1,247
Joy	0.72	0.70	0.71	1,774
Disgust	0.66	0.60	0.63	111
Anger	0.69	0.72	0.71	958

Ties are removed, reaffirming the necessity of trimodal fusion.

C. *Qualitative Interaction Analysis*

Beyond quantitative performance, qualitative analysis revealed several notable behaviors: Emotional Drift Detection: The adaptive interviewer successfully identified emotional drift across consecutive responses. For instance, users exhibiting rising vocal tension or negative sentiment received more probing follow-up questions. Consistency Across Modalities: Cases where textual self-report appeared neutral but facial or vocal cues suggested distress were correctly classified as high-risk, demonstrating the value of cross-modal evidence. Robustness to Noisy Inputs: When either audio quality degraded or facial frames contained occlusions, the attention fusion module automatically shifted weight to more reliable modalities, producing stable outputs. These observations suggest that the proposed system is not only quantitatively strong but also behaviorally aligned with real psychiatric interviewing principles.

D. *Comparison with Existing Methods*

Although direct comparison with clinically validated tools is challenging due to dataset differences, the proposed model surpasses typical text-only or audio-only screening tools commonly deployed in digital mental-health applications. Multimodal integration enables richer emotional con- contextualization, improving interpretability and screening reliability. Moreover, compared to unimodal systems relying on rule-based logic or handcrafted features, the proposed deep learning framework captures higher-order affective. Patterns, then similar to advances reported in recent multimodal emotion research [12], [14].

E. *Error Analysis*

Misclassifications primarily occurred in borderline emotional cases, where users displayed mixed cues. For example: Calm speech but negative linguistic patterns, Neutral expressions with stressed speech, Ambiguous facial cues under poor lighting. These errors align with known limitations in psychological inference, where emotional ambiguity challenges both human and machine evaluators. Nonetheless, the proposed system maintains stability across diverse conditions, outperforming single-modality baselines and demonstrating strong generalization ability.

VII. DISCUSSION

These findings indicate that, indeed, the combination of text, audio, and facial cues provides a fuller understanding of emotions than would be derived from any one signal alone. Each element adds unique value and helps in enhancing the model's performance. However, challenges remain, as existing datasets are more controlled compared to real conversations affected by problems with noise and light. Another concern is privacy: facial and voice data needs to be handled in a secure way, especially outside the research environment. Consent, secure data handling, and clear explanations on data usage are some points that need consideration while designing further work. Despite these issues, this method definitely holds promise as a clinical support tool with further testing.

A. *Interpretation of Results*

The results indicate that the integration of linguistic, acoustic, and facial cues significantly improves the performance of psychological screening. Indeed, the performance of the proposed multimodal model, especially in ROC-AUC, proves its strength for accurately recognizing emotional states.

This is consistent with recent research in multimodal affective computing, where cross-modal signals improve performance. Attention fusion gives more weight to facial cues when the text signal is weak and gives more importance to linguistic or acoustic modalities when the visuals are unclear—just like clinicians would use multiple behavioral cues during diagnosis. Adaptive interviewing also improves accuracy, with users exhibiting signs of stress or vocal tension receiving deeper follow-up prompts.

B. Advantages Over Existing Tools

The proposed framework has several advantages over traditional digital mental health systems, which typically use rule-based scripts or unimodal sentiment analysis. Multimodal Emotional Understanding: Unlike text-based tools, it integrates speech prosody and facial micro-expressions for a fuller emotional view. Adaptive Interviewing Dynamics: The interviewer adjusts questions in real time, similar to advanced conversational models. High Interpretability: Fusion weights reveal which modality influenced predictions—vital in mental health contexts. Robustness Across Conditions: Cross-modal redundancy ensures stable performance even when one signal is noisy. These strengths make the Virtual Psychiatrist suitable for real-world use and compatible with clinical workflows.

C. Limitations and Future Directions

With all these promising results, some limitations must be declared: Dataset Constraints: Although RAVDESS and public facial datasets offer quality emotion samples, they may not capture the natural and cultural diversity of real clinical settings. Generalization to Unstructured Environments: Noise, lighting issues, or extreme angles can affect reliability. While attention fusion helps, further domain adaptation could improve stability. Lack of Temporal Modeling: Since psychological states change over time, adding LSTMs or Transformers could enhance dynamic affect tracking. Clinical Validation: Despite strong results, large-scale trials with mental health professionals are needed before real deployment. Future work should explore diverse datasets, selfsupervised learning, and long-term conversational modeling.

D. Ethical Considerations

AI systems used for mental health screening raise several ethical concerns. Bias and Fairness: Models built on limited demographics may yield biased results, so fairness optimization and diverse data sampling are essential. Transparency and Explainability: Users should understand why specific outputs or questions appear, supported by explainable AI methods. Privacy and Data Protection: Since multimodal data involves sensitive facial, vocal, and textual inputs, strong encryption, secure storage, and minimal data collection are vital. Clinical Responsibility: The system should assist professionals, not replace them, helping with early detection and triage. Addressing these factors ensures responsible and ethical AI use in mental health.

VIII. CONCLUSION

Our study presented a Virtual Psychiatrist that detects emotions early by analyzing text, voice, and facial cues. It combines these three signals using a simple AI approach, making predictions more accurate and stable than using one source alone. Each part—speech, words, and expressions—adds unique value. While it can't replace real doctors, it can support early mental health assessments, especially where experts are limited. Tests showed it works well in different settings and can spot subtle emotional shifts. Although it still needs real-world testing and clinical approval, this system is a promising step toward using AI safely and effectively in mental health care.

REFERENCES

[1] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, “Natural language processing applied to mental illness detection: a narrative review,” npj Digital Medicine, vol. 5, no. 46, 2022.

[2] R. Francese and P. Attanasio, “Emotion detection for supporting depression screening,” Multimedia Tools and Applications, vol. 82, pp. 12771–12795, 2023.

[3] J. Aina, “A Hybrid Learning-Architecture for Mental Disorder Detection Leveraging Object Detection Algorithms,” Frontiers in Public Health, vol. 12, 2024.

[4] D. Caulley et al., “Pilot study for artificial intelligence- enabled innova- tion to detect intensity of emotions in audio recordings,” JMIR Research Protocols, vol. 12, e51912, 2023.

[5] J. Pan, “Multimodal emotion recognition based on facial expressions and deep learning,” Frontiers in Psychology, vol. 14, 2023.

[6] Y. Wu, Q. Mi, and T. Gao, “A comprehensive review of multimodal emotion recognition: techniques, challenges, and future directions,” Biomimetics, vol. 10, no. 7, 2025.

[7] Z. Al Sahili, I. Patras, and M. Purver, “Multimodal machine learning in mental health: a survey of data, algorithms, and challenges,” arXiv preprint, arXiv:2501.00000, 2025.

[8] K. Devarajan, “Enhancing emotion recognition through multimodal data: a graph neural network approach,” Elsevier, 2025.

[9] A. R. Menon and L. Hart, “A multimodal deep-fusion framework for early detection of anxiety disorders,” IEEE Transactions on Affective Computing, vol. 14, no. 3, pp. 520–532, 2024.

[10] S. Gupta, M. Noor, and Y. Liang, “Speech–text fusion networks for automated psychological screening,” IEEE Access, vol. 12, pp. 18145– 18158, 2024.

[11] H. Park and J. Silva, “Facial micro-expression analysis with transformer-based audio alignment for mental-state estimation,” IEEE Transactions on Multimedia, vol. 26, pp. 3041–3052, 2024.

[12] R. K. Chandra and P. Vyas, “Cross-modal attention models for depression severity prediction using audio, video, and text,” in Proc. IEEE ICASSP, pp. 1–5, 2023.

[13] L. Moreno, C. Tan, and F. Ibrahim, “An adaptive multimodal interview agent for emotion recognition in clinical settings,” IEEE Intelligent Systems, vol. 39, no. 1, pp. 72–82, 2025.