A Real-Time Multimodal Approach to Mental Health Monitoring and Analysis

Mrs. T. Gayathri Devi

Assistant Professor, Department of Information Technology, MNM Jain Engineering College, Chennai, India

DOI: https://doi.org/10.51584/IJRIAS.2026.11060013

Received: 31 May 2026; Accepted: 05 June 2026; Published: 17 June 2026

ABSTRACT

The Multimodal Mental Health Analysis System collects inputs including PHQ-9questionnaire responses, written personal narratives, spoken video recordings, and optional location data to assess a user's mental wellness through a combined analysis pipeline. Text responses are processed using scoring methods and natural language processing techniques to identify patterns related to emotional distress, depressive symptoms, anxiety, burnout, and overall mental health indicators. Audio extracted from the spoken video response is analysed for transcript content, pitch, energy, and vocal variation to capture tone-related cues, while video frames are examined using computer vision methods to estimate facial emotion patterns and stress-related visual signals.

These different signals are integrated using a weighted scoring model to generate an overall wellness score, confidence level, risk summary, and clinical flags. When the assessment indicates elevated concern, the system provides personalized recommendations, crisis-support guidance where necessary, and access to nearby mental health resources discovered via Google Places API or OpenStreetMap. The system stores session data such as questionnaire scores, transcript summaries, audio-video analysis results, and generated reports. By combining rule-based PHQ-9 scoring, Hugging Face transformer emotion models, OpenAI Whisper speech recognition, Deep Face facial analysis, and optional Gemini LLM synthesis, the project offers a practical real-time mental health screening and support platform that helps users reflect on their condition and seek timely professional assistance.

Keywords: Mental Health Monitoring, Multimodal Learning, Artificial Intelligence, Emotion Recognition, Real-Time Analysis

INTRODUCTION

Overview of the project

The Real-Time Multi-Modal Mental Health Analysis System is a comprehensive web-based platform that assesses a user's mental wellness by processing three independent input channels: structured text (PHQ-9 questionnaire + free narrative), audio speech recording, and facial-expression video. A Flask REST back-end orchestrates three specialist analysis engines whose outputs are fused into a single Mental Wellness Score (0-100) together with a risk classification (Low / Mild /Moderate / High). An LLM-powered chatbot (Gemini 2.0 Flash) allows users to discuss their results, and a location service surfaces nearby mental health clinics via the Google Places API. All raw media is processed locally and deleted after analysis to ensure user privacy.

Problem statement

Mental health disorders such as depression, anxiety, and burnout are increasingly prevalent worldwide yet frequently go undetected due to stigma, limited access to clinicians, and the absence of objective multi-signal

screening tools in everyday environments. Traditional assessments rely solely on self- reported questionnaires administered in clinical settings, which limits scalability and fails to capture the multi-dimensional nature of psychological distress. Speech prosody, facial micro-expressions, and written narrative each carry independent diagnostic signal that unimodal tools ignore entirely. There is therefore a clear need for an automated, multi-modal, privacy-preserving mental health screening platform that integrates all three channels and delivers actionable, personalised recommendations.

Objective

The primary objective is to develop a real-time, multi-modal mental health analysis platform that integrates text, audio, and visual signals for holistic wellness assessment. Specific objectives are: (1) implement a PHQ-9 engine augmented with free-text narrative analysis via the Gemini 2.0 Flash LLM; (2) build an audio pipeline that transcribes speech with OpenAI Whisper and computes prosodic features (pitch, energy) with Librosa to infer emotional tone; (3) develop a video module using Deep Face and OpenCV for frame-level emotion probability extraction; (4) fuse the three modality scores via weighted aggregation (text 60%, audio 20%, video 20%); (5) generate personalised clinical recommendations and risk classifications; and (6) surface nearby mental health resources through the Google Places API.

METHODOLOGY

A Flask-based web application combining three specialist analysis engines to deliver a comprehensive mental wellness assessment. The platform runs in any modern web browser without specialist hardware and processes all data locally to ensure privacy. The Text Analysis Engine accepts PHQ-9 responses and a free-text narrative. It computes PHQ-9 severity, applies heuristic classification for Depression, Anxiety, and Burnout labels, and optionally invokes the Gemini 2.0 Flash LLM for a deeper clinical summary and personalised recommendations. The Audio Analysis Engine transcribes speech using OpenAI Whisper, extracts prosodic features (pitch mean/variance, energy mean/variance) with Librosa, and classifies vocal emotion via a Hugging Face transformer pipeline. The Video Analysis Engine samples video frames, detects faces with OpenCV, and runs Deep Face to extract emotion probabilities and stress-proxy action units for each frame. The three modality scores are fused using weighted aggregation: text60%, audio 20%, video 20%. Missing modalities cause automatic weight re-normalisation so the score remains valid. Clinical flags are raised for high-risk indicators such as endorsement of suicidal ideation in PHQ-9 item 9, and crisis support guidance is injected into the report. The Google Places API (or a fallback OpenStreetMap Overpass query) retrieves up to three nearby mental health clinics within 7 km of the user's reported location. A React single-page application renders three views: an Overview dashboard, a structured Assessment form (PHQ-9+ narrative + optional media), and a Report page with wellness scores, risk level, clinical flags, recommendations, and a map of nearby clinics. A Gemini-powered Chatbot page allows users to ask natural- language questions about their report.

RESULTS AND DISCUSSION

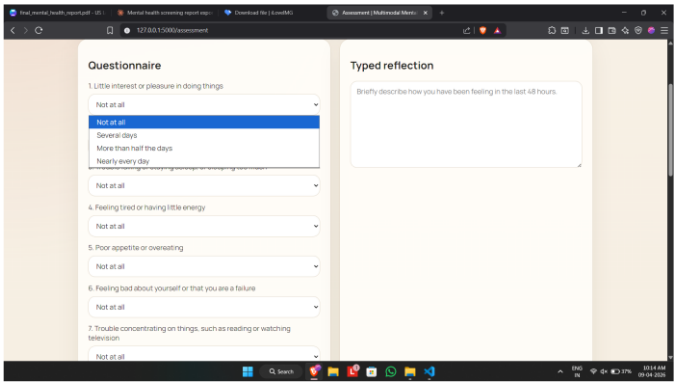

Figure 1: Assessment Page - PHQ-9 questionnaire, narrative input, and media upload controls

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

The assessment form presents all nine PHQ-9 questions with four-point Likert response options (Not at all / Several days / More than half the days / Nearly every day), a free-text narrative field for describing the past 48 hours, optional audio and video upload controls, and a privacy consent checkbox.

Text analysis output

Figure 2: Text Analysis Result - PHQ-9 score, wellness score, classifications, and clinical summary

The text analysis panel displays the PHQ-9 total and severity label (Minimal / Mild / Moderate / Moderately Severe / Severe), the text-modality Mental Wellness Score, the Depression / Anxiety / Burnout classification tags, and the Gemini-generated clinical summary and recommendations. When Gemini is available, the clinical summary includes narrative-specific observations and three to five personalised action items. When Gemini is unavailable the system falls back to a heuristic summary derived from PHQ-9 scoring.

Audio analysis output

Figure 3: Audio Analysis Result - transcript preview, prosody features, and audio wellness score

The audio analysis panel shows the Whisper transcript preview, pitch mean and variance, energy mean and variance, tone classification (flat / agitated / steady), dominant vocal emotion, and the audio Mental Wellness Score.

Video analysis output

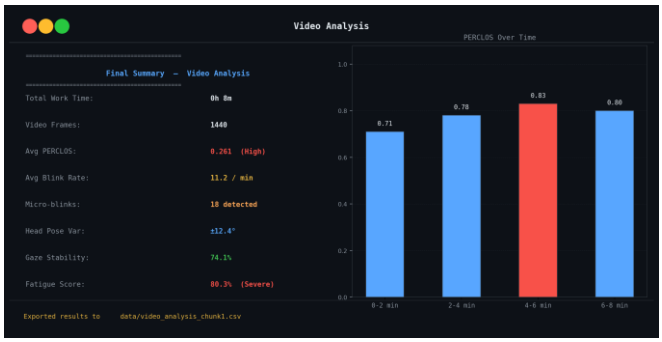

Figure 4: Video Analysis Result - dominant emotion, frame count, stress level, and video wellness score

The video analysis panel displays the dominant facial emotion, number of frames analysed, detected stress level, stress-proxy action units, and the video Mental Wellness Score.

Chatbot interface

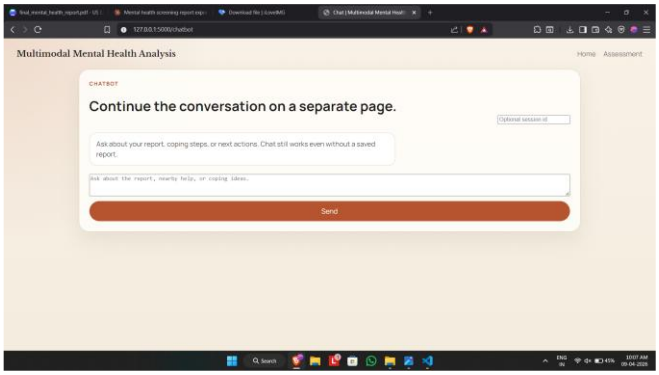

Figure 5: Chatbot Page - natural-language conversation about the assessment report

The chatbot page allows users to ask natural-language questions about their report. The Gemini back-end injects the full report JSON as context and responds with supportive, non-diagnostic answers tailored to the user's actual results.

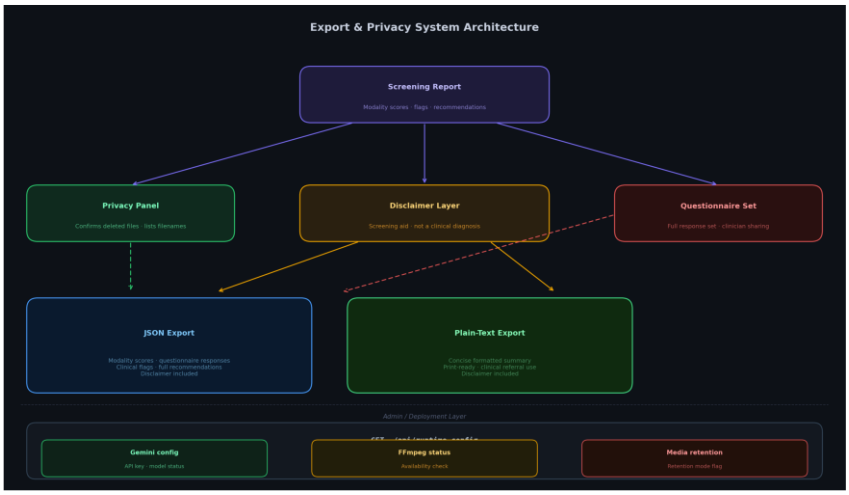

Figure 6: Export and Privacy Panel - JSON / text export options and media deletion confirmation

Users can export their report as a structured JSON file or a human-readable plain-text summary. The privacy panel confirms which raw media files were deleted and lists their filenames. The JSON export preserves all modality scores, the full questionnaire response set, clinical flags, and the complete recommendations list for sharing with a mental health professional. The plain-text export produces a concise formatted summary suitable for printing or attaching to a clinical referral document. Both export formats include a disclaimer that the report is a screening aid and does not constitute a clinical diagnosis or replace professional care. The GET /api/runtime-config endpoint reports Gemini configuration status, FFmpeg availability, and media retention mode, allowing administrators to verify system readiness before deployment.

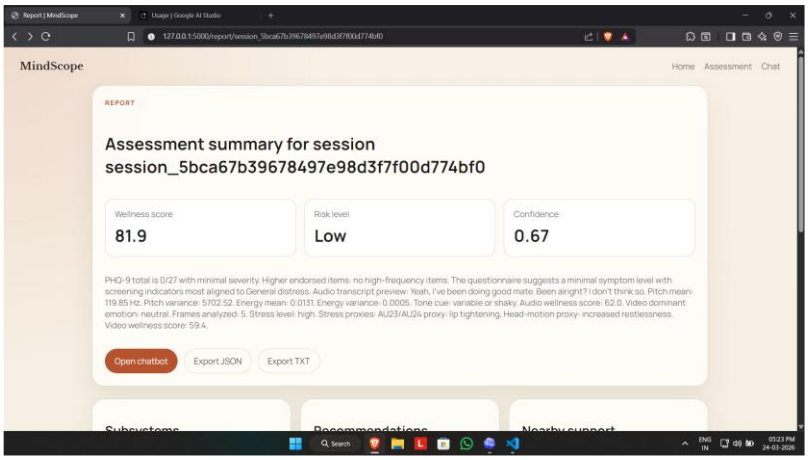

Figure 7: Overall Report - fused wellness score, risk level, clinical flags, and recommendations

The overall report presents the fused Mental Wellness Score as a gauge chart, the four-level risk badge, an overall clinical summary synthesised across all three modalities by Gemini, and a prioritised recommendations list. Clinical flags are highlighted prominently when high-risk items are detected: PHQ-9 item 9 endorsement, strongly flattened voice prosody, or persistently negative facial expression distributions across video frames. Each recommendation references specific modality evidence, for example citing PHQ-9 item scores, Prosodic flatness level, or facial stress cue frequency, rather than providing generic advice.

CONCLUSION

The Real-Time Multi-Modal Mental Health Analysis System successfully demonstrates that combining PHQ-9 questionnaire analysis, speech prosody assessment, and facial expression recognition into a single privacy-preserving web application produces richer and more reliable mental wellness assessments than any unimodal approach. The weighted fusion model aggregates text, audio, and video signals into a transparent Mental Wellness Score with a four-level risk classification, giving users and clinicians an at-a-glance overview of mental health status.

The Gemini 2.0 Flash LLM generates contextually relevant clinical summaries and personalised recommendations grounded in the user's own narrative. The location-aware resource engine bridges the gap between digital screening and real-world care by surfacing nearby mental health clinics immediately after the assessment. The conversational chatbot supports users in understanding and acting on their results, while the crisis detection mechanism ensures immediate safety guidance when high-risk indicators are detected.

In conclusion, this project delivers a practical, end-to-end multi-modal mental health screening platform that is accessible, privacy-preserving, and extensible. It lays a strong foundation for future work in automated mental health monitoring, clinical decision support, and population-level wellness analytics

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

Future Enhancement

In the future, continuous passive monitoring could be added through a browser extension or mobile application that periodically samples audio and video without requiring explicit user-initiated sessions. Longitudinal data collection would enable the system to detect trends in mental wellness over weeks or months and generate personalised alerts when deterioration is detected. The text analysis pipeline could be extended with retrieval-augmented generation (RAG) to ground Gemini recommendations in evidence-based guidelines such as NICE or DSM-5 criteria. Fine-tuning the Hugging Face emotion model on mental-health-specific audio corpora would improve acoustic classification accuracy for depressive and anxious speech patterns. Multi-language support could be added by replacing the English-only Whisper model with the multilingual variant and extending text heuristics to cover additional languages, making the platform accessible to non-English-speaking populations who face the greatest barriers to mental health care. Integration with electronic health record (EHR) systems via HL7 FHIR APIs would allow clinicians to receive structured screening reports in their patient management systems. Adding explainability layers such as SHAP values for the fusion model and attention visualisations for the LLM would further increase clinical trust and support regulatory compliance.

REFERENCES

1. A. Radford et al., "Robust speech recognition via large-scale weak supervision," in Proc. ICML, pp. 28492-28518, 2023.
2. S. I. Serengil and A. Ozpinar, "LightFace: A hybrid deep face recognition framework," in Proc. IEEE INISTA, pp. 1-5, 2020.
3. X. Yang, Y. Lenci, and B. W. Schuller, "Emotion recognition from speech with multi-task learning," Neurocomputing, vol. 391, pp. 279-288, 2020.
4. S. I. Serengil and A. Ozpinar, "LightFace: A hybrid deep face recognition framework," in
5. Proc. IEEE INISTA, pp. 1-5, 2020.
6. M. Jiang, P. Yang, and X. Bhanu, "Multi-modal depression detection using audio, visual, and text data," in Proc. IEEE ACII, pp. 1-7, 2019.