Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

MentalSense AI: Real-Time Mental Stress Detection using Multimodal Behavioral Analysis and Personalized Support

Dr. Sumitha C [GUIDE]
Department of Computer Science and Engineering
G Madegowda Institute of Technology
Mandya, Karnataka, India

Tejas S. R.
Department of Computer Science and Engineering
G. Madegowda Institute of Technology
Mandya, Karnataka, India

Darshan Gowda K. L.
Department of Computer Science and Engineering
G. Madegowda Institute of Technology
Mandya, Karnataka, India

Harshith Gowda M. S.
Department of Computer Science and Engineering
G. Madegowda Institute of Technology
Mandya, Karnataka, India

Mrs. Shwetha L [CO-GUIDE]
Department of Computer Science and Engineering
G Madegowda Institute of Technology
Mandya, Karnataka, India

Nithisha M.
Department of Computer Science and Engineering
G. Madegowda Institute of Technology
Mandya, Karnataka, India

Abstract - Mental stress can affect concentration, communication, decision-making, and well-being. This paper presents MentalSense AI, an integrated application for real-time, stress-related behavioral assessment and personalized wellness support. The system combines webcam-based facial-expression analysis, microphone-based voice analysis, confidence-weighted multimodal fusion, continuous monitoring, assessment history, and guided wellness support. The facial pathway uses MediaPipe-based face localization and a convolutional neural network trained on FER2013-style expression classes. The existing Voice V9 pipeline processes 3.5-second audio windows, extracts Mel-frequency cepstral coefficients (MFCCs), delta-MFCCs, spectral descriptors, and chroma features, applies ANOVA F-test feature selection, and uses Logistic Regression, Linear Support Vector Machine, and Extra Trees classifiers. The application uses a React frontend, Python/FastAPI services, and SQLite persistence. Interface screenshots illustrate the implemented monitoring and assessment workflow. Because the available project materials do not establish a reproducible, end-to-end benchmark, no overall accuracy or latency value is claimed. The paper identifies the evaluation needed to validate classification, fusion, and real-time behavior.

Keywords—mental stress assessment; multimodal analysis; facial expression recognition; voice analysis; confidence-weighted fusion; real-time monitoring; personalized wellness support.

I. INTRODUCTION

Mental stress is a common concern in academic and professional environments, where deadlines, workload, multitasking, and long periods of screen use can affect concentration and everyday well-being. Questionnaires and self-reports remain useful, but they do not inherently provide continuous monitoring. A practical digital support application can help users review observable patterns and access low-intensity wellness guidance.

Facial expressions and voice characteristics provide complementary behavioral signals. A facial-expression model identifies visible expression categories, while voice analysis can capture acoustic characteristics such as spectral shape, chroma, and pitch-related variation. These signals are affected by lighting, background noise, speaking style, and individual differences; they should therefore be treated as model outputs rather than as a clinical diagnosis.

MentalSense AI brings together webcam and microphone acquisition, separate facial and voice pipelines, a fusion stage, real-time dashboard components, history/trend views, and personalized support. The goal is to provide one coherent workflow rather than an isolated classifier. The system is intended for supportive, non-clinical use.

A. Problem Statement

Many stress-related applications focus on a single input modality or stop at classification. Such systems may not provide a continuous view of changing outputs, a clear fusion stage, or a path from assessment to supportive action. The engineering problem addressed here is to integrate facial and voice analysis into a real-time application with understandable outputs, monitoring, and wellness support while clearly communicating uncertainty and limitations.

B. Objectives and Contributions

- Real-time acquisition of facial frames from a webcam and audio signals from a microphone.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

- Facial-expression analysis and voice classification using the project's existing model pipelines.
- Confidence-weighted combination of available modality outputs into a consolidated assessment.
- Continuous visualization of facial output, voice waveform, sound level, pitch, fusion results, and trends.
- Personalized wellness recommendations and conversational support, including guided breathing and mindfulness.

II. RELATED WORK

Zhang et al. proposed a real-time deep-learning framework that combines electrocardiogram (ECG), voice, and facial-expression signals for acute stress detection.

[1]. Their work demonstrates the research value of combining complementary modalities. Its experimental results belong to that study and are not attributed to MentalSense AI.

Facial-expression recognition benchmarks such as FER2013 contain grayscale facial images labelled into seven expression categories.

[2]. These labels describe expressions and do not directly establish psychological stress. MediaPipe offers reusable perception components for real-time face localization.

[3], which can support the input stage of a facial analysis pipeline.

Voice analysis commonly uses acoustic representations such as Mel-frequency cepstral coefficients (MFCCs).

[4]. Support Vector Machines and ensemble classifiers are established approaches for classification tasks.

[5], MentalSense AI combines these methods in its existing voice pipeline and emphasizes integration with a monitoring interface and personalized support.

III. PROPOSED SYSTEM

The proposed system contains an input-acquisition layer, two modality-specific analysis paths, a fusion layer, a monitoring dashboard, and a personalized-support layer. Webcam frames are sent to facial analysis, and microphone audio is processed in fixed-duration windows by the existing Voice V9 pipeline. Available modality outputs are passed to the fusion stage and displayed in the dashboard. Optional text assessment is part of the design, but the deployed connection should be confirmed in the current code before it is described as fully operational.

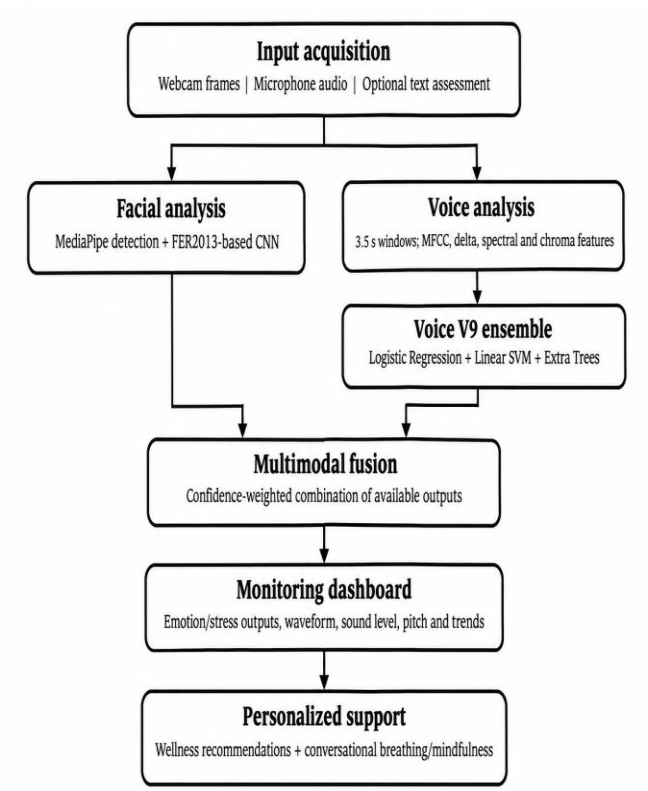


Fig. 1. MentalSense AI architecture: acquisition, facial and voice analysis, fusion, monitoring, and personalized support.

A. Facial Analysis

The facial pathway uses MediaPipe-based face localization to identify the facial region in webcam frames. The region is prepared for the CNN model using grayscale conversion and resizing to 48 × 48 pixels. The expression classes described in the project are angry, disgust, fear, happy, neutral, sad, and surprise. These categories describe visible expressions; any conversion from expression output to a stress-related score requires validation and should not be treated as direct clinical measurement.

B. Voice Analysis

The project's Voice V9 component is retained as the existing, frozen model. Audio is processed in 3.5-second windows. The feature set includes MFCCs, delta-MFCCs, spectral descriptors, and chroma features. Feature selection uses an ANOVA F-test with SelectKBest, and the documented ensemble contains Logistic Regression, Linear Support Vector Machine (SVM), and Extra Trees. Replacing this module or claiming a newer voice version is outside the scope of this paper.

C. Multimodal Fusion

The fusion stage combines modality outputs and their associated confidence values. A conceptual weighted output can be written as $M(t) = \Sigma_i w_i(t)p_i(t)$, where $p_i(t)$ is the output from modality i and $w_i(t)$ is its normalized weight at time t. This equation describes the fusion concept, not a verified exact implementation. The source code should be checked for weight normalization, score calibration, unavailable modalities, and text input handling before final publication.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

D. Monitoring and Personalized Support

The React interface contains separate views for facial analysis, voice waveform, sound level, pitch, and the fused assessment. The project also includes monitoring/history and trend visualizations. Backend services are described as using Python/FastAPI, real-time dashboard communication, and SQLite persistence. Personalized support includes wellness recommendations and conversational guidance such as breathing and mindfulness exercises. These functions are intended to support user awareness and not to diagnose or treat a mental health condition.

IV. IMPLEMENTATION

The frontend is implemented using React and JavaScript/JSX. Python/FastAPI provides backend services, and SQLite is used for application persistence. The realtime workflow initializes the camera and microphone, sends frames to the facial pathway, processes audio windows through Voice V9, combines available outputs, and updates the dashboard. The system also contains user-facing assessment, monitoring, and support flows. Actual behavior of every endpoint and modality connection should be confirmed against the deployed source before claiming full end-to-end operation.

V. RESULTS AND DISCUSSION

The application interface includes monitoring and assessment screens for facial analysis, voice visualization, sound-level and pitch indicators, and fused outputs. The figures provide qualitative evidence of interface organization; they are not a substitute for a controlled model evaluation. The available evaluation record does not provide a complete, reproducible test protocol with independent ground-truth labels for the integrated system, so an overall stress-detection accuracy, F1 score, or latency value is not reported here.

Login Page. The login interface provides the entry point to the MentalSense AI application. It presents the sign-in form and access to the user's stress-support features.

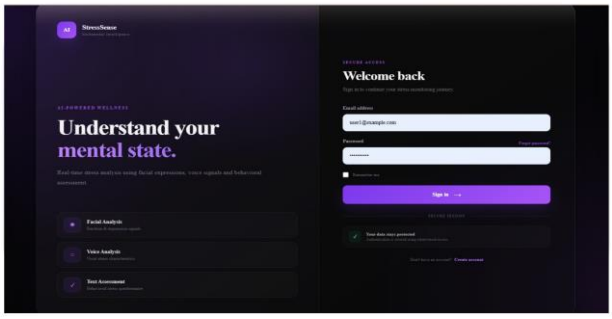


Fig. 2. Login Page.

Stress Assessment Page. The stress-assessment page collects the user's responses to assessment questions. The responses provide a self-report input and should be interpreted separately from the facial and voice model outputs.

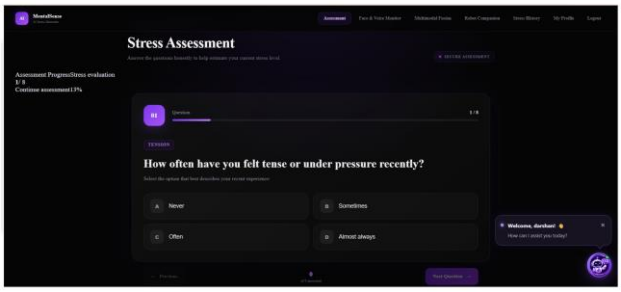


Fig. 3. Stress Assessment Page.

Real-Time Face and Voice Analysis. The real-time monitoring screen shows webcam-based face tracking alongside voice-related visualizations, including waveform, sound level, pitch, and changing analysis outputs.

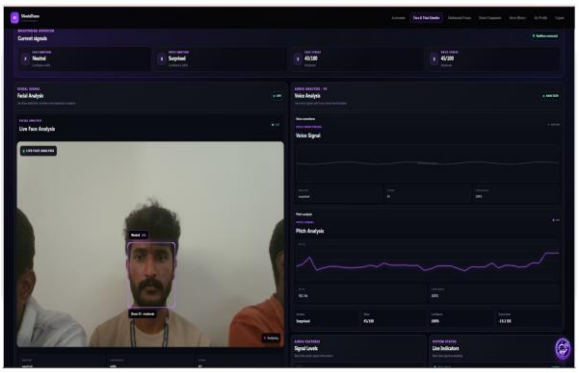


Fig. 4. Real-Time Face and Voice Analysis.

Multimodal Fusion Page. The fusion page brings together the available facial, voice, and other assessment outputs into a consolidated view. The displayed result is a system estimate, not a clinical diagnosis.

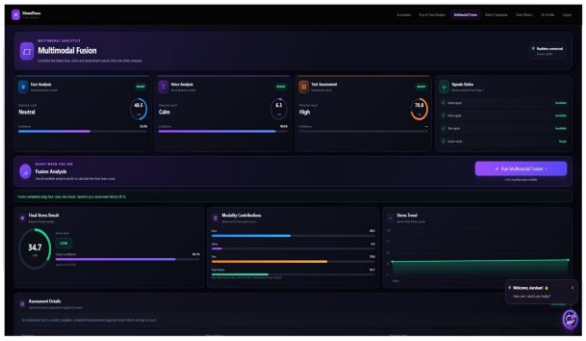


Fig. 5. Multimodal Fusion Page.

Personalized Recommendation Page. The personalized-support screen presents wellness guidance and activities, including breathing or mindfulness support, intended to help users respond constructively to their assessment results.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

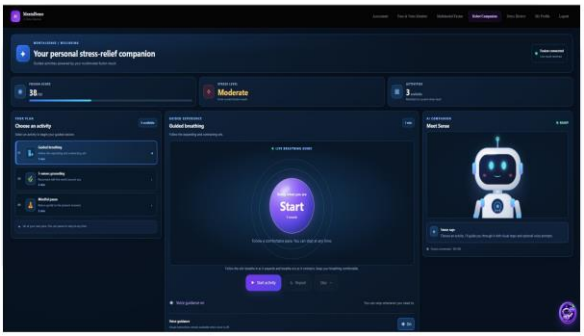


Fig. 6. Personalized Recommendation Page.

Conversational Chatbot Page. The chatbot interface offers a conversational channel for immediate wellness guidance and guided support. It is intended as a supportive tool and not a replacement for professional care.

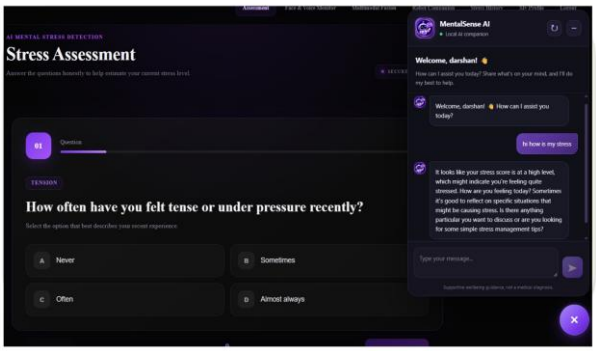


Fig. 7. Conversational Chatbot Page.

Stress History Page. The history page visualizes recorded assessment summaries and a trend chart, helping users review changes over time. Any trend interpretation depends on the quality and consistency of recorded assessments.

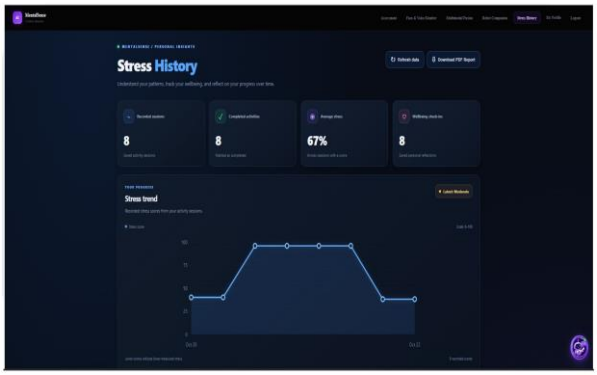


Fig. 8. Stress History Page.

Doctor Portal Page. The doctor dashboard provides a clinician-oriented overview of patient-related records and summary information available in the application. Access and displayed data should follow the project's implemented permissions.

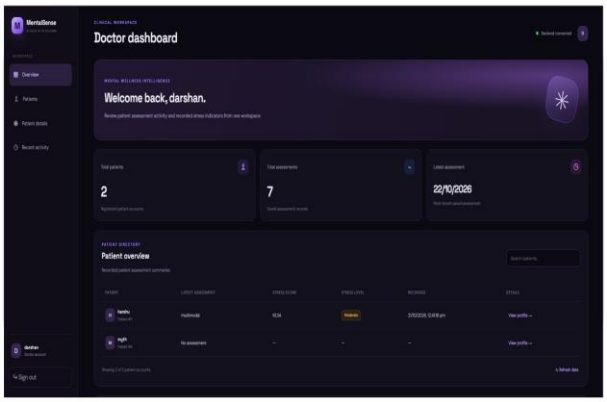


Fig. 9. Doctor Portal Page.

Admin Portal Page. The administration dashboard presents system-level overview information and administrative functions. This screenshot documents the interface; it does not independently verify every backend permission or audit workflow.

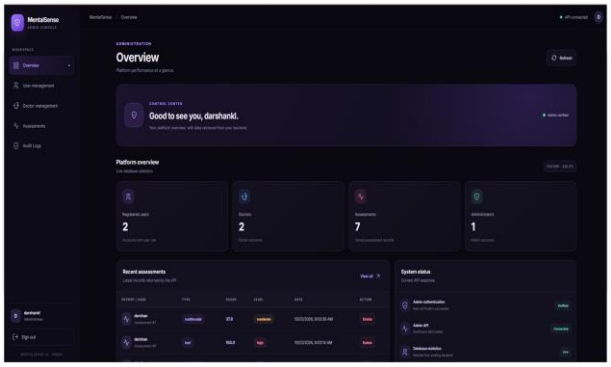


Fig. 10. Admin Portal Page.

A useful evaluation should separate model performance from software integration. For each modality, report the dataset, class distribution, train/validation/test split, preprocessing, confusion matrix, class-wise precision and recall, macro-F1, and accuracy where appropriate. Expression classification results must not be presented as stress-classification results unless a validated stress label protocol is used.

Ablation tests should compare face-only, voice-only, and fused outputs on the same test cases. The evaluation should also test missing or degraded inputs, including poor lighting, background noise, and unavailable microphone/camera permissions. For the realtime application, measure acquisition-to-dashboard latency, audio-window processing time, frame rate, dropped updates, and resource usage on specified hardware.

A. Application Behaviour and Interface

Component	Implementation / role
Frontend	React.js and JavaScript/JSX
Backend	Python and FastAPI services
Data persistence	SQLite database

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

Facial analysis	MediaPipe face localization + CNN expression classification
Voice analysis	Frozen Voice V9; MFCC, delta-MFCC, spectral/chroma features; ANOVA F-test feature selection; Logistic Regression, Linear SVM, Extra Trees
Fusion and support	Confidence-weighted combination of available outputs; monitoring, trends, wellness recommendations, conversational breathing/mindfulness support

VI. LIMITATIONS AND RESPONSIBLE USE

The main limitation is that visible facial expressions and acoustic voice features are indirect behavioral signals and do not independently establish psychological stress. Results may vary with lighting, background noise, language, accent, speaking style, and device quality. The fusion implementation and text pathway should be verified against the current code, and the benefit of fusion should be measured rather than assumed.

Camera and microphone data can contain sensitive personal information. A deployed system should explain the collection purpose, retention period, and access controls; collect only necessary data; and obtain appropriate consent. The system should communicate when a modality is missing and avoid presenting incomplete results as fully multimodal. It must not be represented as a medical diagnostic tool or a replacement for qualified professional care.

VII. CONCLUSION

MentalSense AI integrates facial-expression analysis, voice analysis, confidence-weighted fusion, realtime visualization, monitoring/history, and personalized wellness support into one application workflow. The facial pipeline uses MediaPipe-based localization and a CNN expression model, while the existing Voice V9 pipeline uses acoustic features, feature selection, and an ensemble classifier. The interface connects analysis outputs with trends and supportive guidance.

Future work should verify the deployed fusion pathway, evaluate the text input path, and run controlled experiments with documented datasets and independent test splits. Additional work includes face-only/voice-only/fusion comparisons, real-time latency measurement, usability evaluation, robustness tests across environments, and careful privacy review. Quantitative performance claims should be added only after the experiments are completed and reported transparently.

ACKNOWLEDGMENT

The authors thank the Department of Computer Science and Engineering, G. Madegowda Institute of Technology (GMIT), Bharathinagara, Karnataka, India, for academic support and guidance. Confirm the final corresponding-author email and supervisor wording before submission.

REFERENCES

[1] J. Zhang, H. Yin, J. Zhang, G. Yang, J. Qin, and L. He, “Real-time mental stress detection using multimodality expressions with a deep learning framework,” Frontiers in Neuroscience, vol. 16, Art. no. 947168, Aug. 2022, doi: 10.3389/fnins.2022.947168.

[2] I. J. Goodfellow et al., “Challenges in representation learning: A report on three machine learning contests,” Neural Networks, vol. 64, pp. 59–63, 2015.

[3] C. Lugaresi et al., “MediaPipe: A framework for building perception pipelines,” arXiv:1906.08172, 2019.

[4] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[5] C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, pp. 273–297, 1995.

[6] T. G. Dietterich, “Ensemble methods in machine learning,” in Multiple Classifier Systems, Lecture Notes in Computer Science, vol. 1857, pp. 1–15, 2000.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)