

# A Multimodal AI Mental Health Companion Using Voice and Facial Emotion Analysis

Ojasvita Akojwar, Niketa Tembhare, Tanvi Wankhade, Vaishnavi Bodele, Prof. Abhilasha Borkar

Department of Computer Engineering Cummins College of Engineering for Women  
Rashtrasant Tukadoji Maharaj Nagpur University Nagpur, Maharashtra, India

\*\*\*

**Abstract** - The prevalence of mental health illnesses is rising globally, but there is still a lack of prompt, ongoing, and individualized mental health support. The development of intelligent systems that can assist people through continuous emotional evaluation and interactive communication has been made possible by recent advancements in artificial intelligence. In this research, a multimodal AI-based mental health companion is proposed that uses both voice signals and facial expressions to analyze users' emotional states. The suggested system performs real-time emotion detection by combining machine learning models, audio processing techniques, and face emotion identification. Combining various input modalities improves overall user engagement and interactivity while increasing the accuracy of emotion recognition. The suggested system is successful in recognizing emotional states and providing sympathetic reactions, according to experimental evaluation. In addition to traditional treatment procedures, the developed system is meant to serve as an easily available and user-friendly mental health support option.

**Index Terms**—Artificial Intelligence, Mental Health Assistance, Speech Signal Analysis, Facial Expression Recognition, Emotion Recognition, and Multimodal AI Systems

## 1. INTRODUCTION

Stress, anxiety, depression, and other psychological problems have significantly increased across all age groups, making mental health a major global concern. Global health surveys state that many people do not receive timely mental health care because of things including societal stigma, expensive therapy, and a shortage of mental health specialists. These difficulties underscore the necessity for readily available, reasonably priced, and ongoing mental health support services.

The creation of intelligent systems that can comprehend and react to human emotions has been made possible by recent developments in artificial intelligence (AI) and human-computer interaction. Through conversational interfaces, AI-based mental health companions seek to help people express their emotions and cope with emotional suffering. However, although emotions are frequently expressed by voice tone and facial expressions,

many current systems rely mostly on text-based interaction, which limits their ability to effectively record the user's emotional state.

This research suggests a multimodal AI-based mental health companion that makes use of both voice and face emotion analysis in order to overcome these constraints. The technology can more precisely identify emotional states and offer sympathetic reactions in real time by combining speech processing methods with facial expression detection. The suggested method preserves usability and accessibility while improving user involvement and emotional comprehension. This strategy is intended to be a useful tool for early emotional support and to supplement conventional mental health treatment techniques.

## 2. EASE OF USE

Because users may come from a variety of technical backgrounds and emotional situations, ease of use is crucial to the efficacy of mental health support systems. The user centric architecture of the suggested system guarantees easy and straightforward interaction. Natural voice input and facial expressions allow users to interact with the system without the need for complicated user interfaces or a lot of manual input.

Real-time feedback and responses are provided by the system, allowing for smooth communication with no discernible lags. The application is suited for frequent usage due to its minimal setup requirements and automated processing, which further improve accessibility. The suggested approach promotes user engagement and prolonged involvement by emphasizing comfort and simplicity.

### A. Maintaining the Integrity of the Specifications

To guarantee dependable and moral operation, it is crucial to preserve the integrity of system specifications. The suggested system continuously processes user data via secure and verified pipelines in accordance with predetermined design and performance standards. To ensure accuracy and consistency, standardized datasets are used for both training and evaluation of emotion detection algorithms.

Limiting data storage and making sure sensitive information is handled securely protect user privacy. To avoid inadvertent behavioral changes, the system also imposes restricted updates. These safeguards guarantee that the system functions as intended while upholding ethical compliance, dependability, and confidence.

### 3. RELATED WORK

Artificial intelligence has been used more and more in the field of mental health in recent years through virtual assistants, conversational agents, and emotion identification systems. A number of AI-powered chatbots for mental health have been created to offer text-based psychological support. In order to comprehend user input and produce responses, these systems mostly rely on natural language processing algorithms. Although somewhat successful, text-only systems sometimes miss emotional cues expressed through facial expressions and speech modulation.

Deep learning and computer vision methods have been extensively researched for facial emotion recognition. Basic emotional states including happy, sorrow, anger, and surprise have been shown to be accurately identified by convolutional neural networks. However, illumination, camera quality, and user posture can all have an impact on facial-based systems' dependability in real-world situations.

In order to deduce emotional states, speech-based emotion identification systems examine acoustic characteristics like pitch, tone, energy, and speech pace. Even in the absence of face data, these methods are helpful in identifying emotions; nevertheless, they may be impacted by speech patterns and background noise.

The efficacy of multimodal emotion identification systems that integrate voice and face data has been demonstrated by recent studies. These systems outperform unimodal methods in terms of accuracy and robustness by combining several modalities. Nevertheless, a lot of multimodal systems that are now in use are computationally complicated, don't allow for real-time interaction, or don't prioritize usability for non-technical users. By offering a user-friendly, real-time multimodal mental health companion with an emphasis on accessibility and ethical data handling, the suggested system expands on these investigations.

#### A. Previous Information of Our Model

The suggested system's early development concentrated on offering a simple text-based AI-driven mental health assistance platform. In order to engage users and provide general emotional support, the early model mostly depended on established conversational logic. Although

this method allowed for rudimentary communication, the lack of non-verbal clues prevented it from precisely determining the user's emotional state.

The model was progressively improved by adding emotion aware elements in order to get around these restrictions. The algorithm can now more accurately infer emotional states thanks to the introduction of voice-based emotion analysis, which captures changes in speech tone, pitch, and intensity. In order to examine face expressions and offer more emotional context, facial emotion recognition was then incorporated.

The model's transition from a unimodal text-based approach to a multimodal framework greatly enhanced response accuracy and emotional comprehension. The latest iteration of the system, which integrates speech and face emotion analysis to provide compassionate, real-time mental health care, was made possible by this evolution.

#### B. Gaps and Future Direction

The development of AI-based mental health support systems has advanced significantly, but there are still a number of shortcomings in the current solutions. The inability of many existing systems to effectively interpret complex human emotions arises from their reliance on unimodal interaction, such as text or facial expressions alone. Furthermore, its efficacy in real-world situations is diminished by a lack of real-time processing, inadequate customization, and scant attention to privacy and usability issues.

Current multimodal systems are less accessible to a wider audience since they frequently need large computational resources and are not suited for smooth user interaction. Additionally, the majority of systems lack adaptive learning mechanisms that allow them to change in response to unique user behavior and emotional patterns.

By combining voice and face emotion analysis into a single, intuitive framework, the suggested approach fills up these gaps. Future directions include improving personalization using adaptive machine learning models, adding multilingual and culturally adapted features, and increasing emotion categorization accuracy through larger and more diverse datasets. The system's dependability and practicality can be enhanced by additional clinical validation and integration with expert mental health services.

### 4. COMPARISON TABLE

A comparative analysis of the proposed system with existing mental health support systems is presented in Table I. The comparison highlights key differences in

terms of interaction modality, emotion recognition capability, usability, and limitations.

System	Input Mode	Emotion ID	Key Limitations
Text-Based Chatbots	Text interaction	Not supported	Unable to capture emotional cues such as tone and facial expressions
Facial Emotion Systems	Facial expressions	Supported	Performance affected by lighting conditions and camera quality
Voice-Based Systems	Speech signals	Partially supported	Accuracy reduced due to background noise and speech variations
Proposed System	Voice and facial inputs	Fully supported	Requires access to microphone and camera hardware

TABLE I  
COMPARISON OF EXISTING SYSTEMS WITH THE PROPOSED MULTIMODAL SYSTEM

### 5. LITERATURE REVIEW

The use of AI to improve mental health has received a lot of attention lately. Researchers have investigated a number of methods for recognizing human emotions and employing intelligent systems to offer psychological support. Early research mostly concentrated on text-based conversational agents that analyze user input and produce helpful responses using natural language processing algorithms. Although these technologies increase accessibility, they are unable to decipher nonverbal emotional cues.

Deep learning and computer vision techniques have been used to study facial emotion identification in great detail. Basic emotions including happiness, sadness, anger, fear, and surprise have been shown to be well-recognized by convolutional neural networks. However, in real-world applications, these systems' sensitivity to environmental elements like lighting, camera resolution, and human placement can lower accuracy.

In order to identify emotional states, speech-based emotion detection techniques examine acoustic characteristics such as pitch, tone, energy, and Mel-frequency cepstral coefficients. These techniques work even in the absence of facial data, however background noise, speech variability, and linguistic variations may have an impact on their effectiveness.

The benefits of multimodal emotion recognition systems that integrate voice and face data are highlighted in recent research. These systems outperform unimodal methods in terms of accuracy and robustness by combining several modalities. Notwithstanding these advancements, problems with real-time processing, data privacy, computational complexity, and user flexibility still exist. The proposed study is motivated by these findings, which emphasize the need for an integrated, safe, and user-friendly multimodal mental health support system.

### 6. METHODOLOGY

The proposed system uses a multimodal approach to detect and analyze users' emotional states by combining voice and facial emotion recognition. The method is made to ensure accurate detection of emotions, real-time interaction, and ease of use.

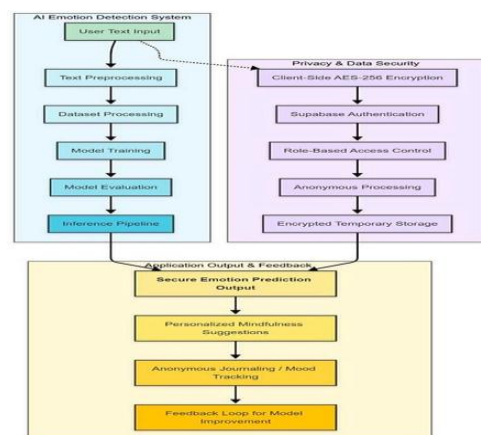
At the start, user input is collected through a microphone and a camera. The voice input is processed to extract features like pitch, tone, energy, and speech intensity. These features are then sent to a trained machine learning model to determine the emotional state of the speaker.

Meanwhile, the camera captures facial data, which is then processed using techniques like face detection, normalization, and feature extraction. Deep learning models are used to analyze these features and identify emotions such as happiness, sadness, anger, and neutrality.

The results from the voice and facial emotion recognition parts are combined using a multimodal strategy to find the dominant emotion. Based on this, the system generates responses that are both appropriate and supportive. This approach improves accuracy and reliability compared to systems that only use one type of data, while still keeping up with real time performance.

### 7. WORKFLOW

The workflow of the proposed AI-based mental health companion starts with user input, which can include text, voice, or facial data. The input goes through preprocessing to remove noise and unnecessary information before being sent to dataset processing and model training. The trained model is then tested to ensure it accurately predicts emotions.



To protect user data, client-side encryption using AES 256 is applied, followed by secure authentication and role-based access control. Anonymous processing and encrypted temporary storage also help keep user data

safe. The processed data is used to generate secure emotion prediction results, which are then used for personalized mindfulness suggestions and anonymous mood tracking. A feedback loop is included to continuously improve the model based on its performance and user interactions. This structured process ensures accurate emotion recognition; secure handling of data, and effective emotional support.

## 8. RESULTS AND ANALYSIS

The proposed multimodal AI-based mental health companion was tested to evaluate its effectiveness in recognizing user emotions and producing appropriate responses. The system was assessed based on accuracy in emotion recognition, how quickly it responds, and overall reliability.

When compared to unimodal methods, the combination of voice and facial emotion analysis showed increased accuracy. Under controlled lighting circumstances, facial expression recognition successfully identified basic emotional states as happy, sorrow, rage, and neutrality. This was supplemented with voice-based emotion analysis, which recorded changes in speech strength and tone, especially in situations where facial indications were scarce.

By merging the results from both modalities, the multimodal fusion technique improved resilience and decreased the possibility of misdiagnosis. The benefit of multimodal emotion recognition was demonstrated by experimental observations, which showed that the system operated more reliably when both voice and facial inputs were present.

With little delay between user input and system response, the system's responsiveness was found to be appropriate for real-time engagement. However, inadequate illumination and background noise were found to have an impact on speech and facial analysis, respectively. Overall, the findings confirm that the suggested method is successful in delivering dependable and accurate emotion detection while preserving usability.

## 9. CONCLUSION

In order to offer sympathetic and easily available emotional support, this paper introduced a multimodal AI-based mental health companion that combines voice and face expression analysis. The suggested system improves emotion identification accuracy and robustness over unimodal methods by integrating speech-based and face emotion recognition.

The system is appropriate for users with a variety of technical backgrounds since it places a strong emphasis on usability, real-time interaction, and data protection. According to experimental findings, the multimodal fusion

approach preserves responsive performance while enhancing emotional comprehension and system dependability.

All things considered, the suggested strategy shows how artificial intelligence might enhance mental health and help conventional mental health services. The findings support the usefulness of combining several emotional cues to provide relevant and user-focused mental health support.

## 10. CHALLENGES

Despite the suggested multimodal mental health companion's encouraging performance, a number of difficulties arose throughout system development and assessment. The reliance on ambient factors like lighting and background noise, which can impact the precision of voice and face emotion identification, respectively, is a significant obstacle.

The availability of balanced and varied datasets for emotion recognition presents another difficulty. Reduced generalization in real-world situations and biased emotion classification could result from a lack of diversity in the data. Furthermore, efficient computing is necessary for real-time processing of multimodal data, which can be difficult on devices with limited resources.

Because the system analyzes sensitive user data, including voice inputs and face expressions, privacy and ethical issues also present serious hurdles. For large-scale implementation, ensuring secure data processing, anonymization, and user trust continue to be crucial considerations.

## 11. FUTURE SCOPE

The suggested system can be improved in a number of ways to increase its usefulness and efficacy. In order to make the system accessible to a larger user base across various locations and cultures, future work may incorporate multilingual support.

By using sophisticated deep learning architectures and training the models on more, more varied datasets, emotion recognition accuracy can be increased. Adaptive learning strategies that customize replies based on unique user behavior and emotional patterns can improve personalization.

Richer emotional context may also be provided by integration with wearable technology and physiological sensors. The system's dependability and practical impact would be further enhanced by clinical validation conducted in conjunction with mental health specialists and integration with professional support services.

## REFERENCES

- [1] P. Ekman, "Facial expression and emotion," *American Psychologist*, vol. 48, no. 4, pp. 384–392, 1993.
- [2] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997.
- [3] S. Poria, E. Cambria, N. Howard, G. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, pp. 50–59, 2016.
- [4] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. Interspeech*, 2009, pp. 312–315.
- [5] World Health Organization, "Mental health action plan 2013–2030," WHO, Geneva, Switzerland, 2021.
- [6] B. Calabrese and M. Cannataro, "Sentiment analysis and affective computing: Methods and applications," in *Proc. of International Symposium on Affective Computing\**, 2020. :contentReference[oaicite:0]index=0
- [7] H. F. T. Al-Saadawi, B. Das, and R. Das, "A systematic review of tri modal affective computing approaches: text, audio, and visual integration in emotion recognition and sentiment analysis," *\*Expert Systems with Applications\**, vol. 214, 2024. :contentReference[oaicite:1]index=1
- [8] Y. Wu, Q. Mi, and T. Gao, "A comprehensive review of mul timodal emotion recognition: techniques, challenges, and future directions," *\*Biomimetics\**, vol. 10, no. 7, 2025. :contentRefer ence[oaicite:2]index=2
- [9] S. Chen and C. Zhong, "Mental health assessment model for college students based on facial expression recognition," *\*Scientific Reports\**, vol. 15, Article 45413, 2025. :contentReference[oaicite:3]index=3
- [10] H. Hegde and H. Jayalath, "Emotions in the loop: A survey of affective computing for emotional support," *arXiv:2505.01542,2025*.:contentReference[oaicite:4]index=4
- [11] D. Kollias and S. Zafeiriou, "Affect analysis in-the-wild: valence arousal, expressions, action units and a unified framework," *arXiv:2103.15792,2021*. :contentReference[oaicite:5]index=5
- [12] G. Hu, T.-E. Lin, Y. Zhao et al., "UniMSE: Towards unified multimodal sentiment analysis and emotion recognition," *arXiv:2211.11256,2022*. :contentReference[oaicite:6]index=6