

An AI-Driven Multimodal Framework for Depression Screening Using Conversational Analysis and Facial Emotion Recognition

¹Dr.A.Nagarjuna Reddy,

Professor,

Department of Computer Science and Engineering, Sridevi Women's Engineering College,
Wipro V.N.Pally, Gachibowli, Hyderabad West IT Corridor, R.R.District-500 075

²Dr.P.Suneel Kumar,

Professor,

Department of Electronics and Communication Engineering, Sridevi Women's Engineering College,
Wipro V.N.Pally, Gachibowli, Hyderabad West IT Corridor, R.R.District-500 075

³Dr. P.Avinash,

Professor,

Department of Information Technology, Sridevi Women's Engineering College, Wipro V.N.Pally,
Gachibowli, Hyderabad West IT Corridor, R.R.District-500 075

⁴Dr Md Nazmoddin,

Associate Professor,

Department of Computer Science & Engineering, Sridevi Women's Engineering College,
Wipro V.N.Pally, Gachibowli, Hyderabad West IT Corridor, R.R.District-500 075

Abstract

The proposed **RT-DepDetect** algorithm introduces a multimodal AI framework for real-time depression detection by integrating chatbot-based conversational analysis with live facial emotion recognition. Initially, user interaction is captured through an intelligent chatbot that processes textual responses using natural language processing (NLP) techniques. Sentiment analysis and contextual embedding models are applied to extract emotional indicators from user conversations. Simultaneously, live video input is analyzed using a deep convolutional neural network to detect facial expressions such as sadness, stress, and neutral states. A temporal feature extraction module captures micro-expressions and emotional variations over time. The textual and visual features are fused using an attention-based multimodal fusion layer, assigning adaptive weights to each modality based on confidence levels. A classification module then predicts depression severity levels using a softmax-based decision layer. An iterative feedback mechanism continuously updates predictions based on ongoing interaction, improving accuracy and personalization. The proposed algorithm enhances early detection, ensures real-time monitoring, and provides a scalable and non-invasive solution for mental health assessment.

Keywords

Depression Detection, Artificial Intelligence, Chatbot Systems, Natural Language Processing (NLP), Facial Emotion Recognition, Multimodal Learning, Deep Learning, Convolutional Neural Networks (CNN), Sentiment Analysis, Real-Time Monitoring, Mental Health Assessment, Human-

Computer Interaction, Emotion Classification, Feature Fusion, Attention Mechanism, Behavioral Analysis, Computer Vision, Healthcare AI, Affective Computing, Intelligent Diagnostic Systems

Introduction

Depression is a major mental health disorder affecting millions worldwide [1]. Artificial intelligence has improved mental health diagnosis through automated systems [2]. Chatbot-based systems enable interactive mental health assessment [3]. Natural language processing enhances emotional understanding from text [4]. Deep learning models improve classification accuracy in healthcare applications [5]. Facial emotion recognition provides real-time behavioral insights [6]. Computer vision techniques detect subtle facial expressions [7]. Multimodal systems combine text and visual data for better prediction [8]. Attention mechanisms enhance feature fusion efficiency [9]. Real-time monitoring improves early detection [10]. AI-driven healthcare systems support scalable diagnosis [11]. Human-computer interaction enhances user engagement [12]. Behavioral analysis improves mental health assessment [13]. Emotional analytics support decision-making [14]. Hybrid AI models improve robustness [15]. Data-driven approaches enhance prediction accuracy [16]. Adaptive systems enable personalized monitoring [17]. Continuous learning improves model performance [18]. Integrated frameworks enhance reliability [19]. However, challenges remain, motivating advanced multimodal AI-based depression detection systems [20].

Motivation

10.48047/jocaaa.2026.35.07.19

Recent literature highlights significant advancements in depression detection using AI, chatbots, and facial emotion recognition techniques. Natural language processing enables effective analysis of user conversations, while computer vision models capture emotional cues from facial expressions. Multimodal approaches combining text and visual data have improved detection accuracy and robustness. However, existing systems often operate independently on single modalities, limiting overall performance. Additionally, many approaches lack real-time adaptability and personalized assessment capabilities. Motivated by these gaps, the proposed work aims to develop a unified multimodal framework that integrates chatbot-based interaction and live facial analysis to enhance accuracy, enable real-time monitoring, and support early depression detection.

Challenges

AI-based depression detection systems face several critical challenges. One major issue is limited availability of labeled mental health datasets, affecting model training and generalization [1]. Accurate facial emotion recognition in real-world conditions is difficult due to lighting variations, occlusions, and diverse expressions [2]. Additionally, contextual understanding of user conversations remains challenging for NLP models, leading to misinterpretation of emotional states [3]. Multimodal fusion introduces high computational complexity, limiting real-time deployment [4]. Furthermore, ensuring privacy, security, and ethical handling of sensitive mental health data is a significant concern for practical adoption [5].

Problem Statement

Despite advancements in AI-based depression detection, significant gaps remain in existing approaches. Current systems often rely on single modalities such as text or facial analysis, limiting detection accuracy and robustness [1]. Multimodal methods improve performance but lack efficient integration and real-time adaptability [2]. Additionally, existing models struggle with contextual understanding of user emotions and variability in facial expressions under real-world conditions [3]. High computational complexity further restricts real-time deployment [4]. Moreover, privacy concerns and lack of personalized assessment reduce practical usability [5]. Therefore, a unified, efficient, and adaptive multimodal framework is required for accurate and real-time depression detection.

Importance of the Proposed Solution

The proposed multimodal AI-based depression detection system is essential for improving early diagnosis and continuous mental health monitoring [1]. By integrating chatbot-based conversational analysis with real-time facial

emotion recognition, the system enhances detection accuracy and robustness [2]. The unified framework enables real-time assessment and personalized monitoring, addressing limitations of existing single-modality approaches [3]. Additionally, the proposed solution is designed to optimize computational efficiency, making it suitable for scalable deployment [4]. Improved accuracy and early detection can support timely intervention, reduce mental health risks, and enhance overall healthcare outcomes in AI-driven mental health systems [5].

Objectives

The primary objective of this work is to design a novel multimodal AI-based algorithm for real-time depression detection using chatbot interaction and facial emotion recognition [1]. The algorithm aims to develop an effective natural language processing module for accurate sentiment and context analysis from user conversations [2]. It focuses on integrating deep learning-based facial expression recognition to capture real-time emotional cues [3]. Additionally, the design seeks to implement an efficient multimodal fusion strategy for improved prediction accuracy [4]. Another objective is to ensure low computational complexity and enable personalized, real-time mental health monitoring for practical deployment [5].

Overview of the Paper

This paper presents a multimodal AI-based framework for real-time depression detection using chatbot interaction and facial emotion recognition [1]. Section 2 reviews existing literature and identifies limitations in current approaches [2]. Section 3 describes the system model and problem formulation, including text and facial feature extraction [3]. Section 4 introduces the proposed algorithm, detailing multimodal fusion and classification mechanisms [4]. Section 5 presents experimental results and performance evaluation using metrics such as accuracy and precision [5]. Finally, Section 6 concludes the paper and discusses future research directions for improving scalability and real-time deployment.

2. Literature Survey

2.1 AI-Based Depression Detection

Depression detection using artificial intelligence (AI) has gained significant attention due to advancements in machine learning and deep learning techniques for mental health analysis [1]–[5]. AI-based systems enable automated identification of depressive symptoms by analyzing behavioral, textual, and emotional patterns. Natural language processing (NLP) plays a crucial role in extracting meaningful insights from user-generated text such as social media posts, clinical reports, and conversational data [6], [7]. Machine learning models such as support vector

10.48047/jocaaa.2026.35.07.19

machines, decision trees, and neural networks have been widely applied for classification tasks, while deep learning models improve feature representation and detection accuracy [8], [9]. These models identify linguistic patterns such as negative sentiment, reduced engagement, and emotional instability, which are indicative of depression [10]. Furthermore, AI-based systems provide scalable and cost-effective alternatives to traditional clinical assessments, enabling early diagnosis and intervention. However, challenges such as data imbalance, lack of interpretability, and privacy concerns remain significant barriers [11], [12]. Despite these limitations, AI-driven depression detection frameworks continue to evolve, offering promising solutions for real-time mental health monitoring and personalized healthcare systems [13].

2.2 Chatbot and NLP-Based Approaches

Chatbot-based systems have emerged as effective tools for automated depression assessment, providing interactive and continuous mental health monitoring [14]–[18]. These systems simulate human-like conversations and analyze user responses using advanced NLP techniques to detect emotional and psychological states. Deep learning models such as recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and transformer-based architectures significantly enhance sentiment classification and contextual understanding [6], [19]. Chatbots can capture temporal variations in user emotions, enabling dynamic analysis of mental health conditions [20]. Additionally, these systems improve accessibility by offering 24/7 support without requiring direct clinical supervision. Personalized conversational models further enhance user engagement and detection accuracy [21]. However, challenges such as ambiguity in language interpretation, limited emotional depth understanding, and dependency on high-quality training data affect performance [22]. Despite these challenges, integrating NLP with conversational AI systems provides a scalable and efficient solution for depression detection, making chatbot-based approaches a key component in modern mental healthcare systems [23].

2.3 Facial Emotion Recognition Techniques

Facial emotion recognition (FER) techniques play a critical role in depression detection by analyzing visual cues and behavioral patterns [24]–[25]. Convolutional neural networks (CNNs) are widely used for extracting facial features and identifying emotional expressions such as sadness, fatigue, and lack of engagement [5], [8]. These systems process facial images or video sequences to detect subtle changes in facial muscles, eye movement, and expressions that indicate emotional states [9]. Advanced computer vision techniques, including facial landmark detection and deep feature extraction, improve model accuracy and robustness [10], [11]. FER systems enable

real-time monitoring of user behavior, making them suitable for continuous mental health assessment [12]. However, challenges such as varying illumination, occlusions, and differences in individual facial expressions affect system performance [13]. Additionally, cultural variations and dataset biases limit generalization capabilities [14]. Despite these limitations, FER-based approaches provide objective and non-invasive insights into emotional states, significantly contributing to depression detection frameworks [15].

2.4 Multimodal Learning and Integration

Multimodal learning approaches have significantly improved depression detection by integrating multiple data sources such as text, facial expressions, speech, and physiological signals [16]–[20]. These systems leverage complementary information from different modalities to achieve more accurate and robust predictions. Feature fusion techniques, including early fusion and late fusion strategies, enable effective integration of heterogeneous data [21]. Attention mechanisms further enhance model performance by assigning importance to relevant features across modalities [22]. Advanced deep learning architectures, such as multimodal transformers and hybrid CNN-RNN models, facilitate efficient feature extraction and integration [23]. These approaches overcome the limitations of single-modality systems by capturing both behavioral and emotional characteristics [24]. However, multimodal systems introduce challenges such as increased computational complexity, synchronization of data streams, and difficulty in feature alignment [25]. Despite these challenges, multimodal frameworks provide a comprehensive and reliable approach for depression detection, making them a promising direction for future AI-based mental health systems.

2.5 Research Gaps in Existing Methods

Despite significant advancements in AI-based depression detection systems, several research gaps remain that hinder their practical deployment [1], [6], [14], [20], [25]. One major limitation is the lack of real-time adaptability, as many models are trained on static datasets and fail to perform effectively in dynamic environments. Multimodal integration remains complex due to issues such as data synchronization, feature fusion complexity, and increased computational overhead [16], [21]. Additionally, scalability is a concern, as existing systems struggle to handle large-scale real-world data efficiently [18]. Privacy and ethical concerns related to sensitive mental health data further limit practical adoption [12]. Another critical challenge is the lack of interpretability in deep learning models, which reduces trust and transparency in clinical applications [10]. Furthermore, variations in user behavior and cultural differences affect model generalization. These challenges highlight the need for developing adaptive, scalable, and

interpretable AI-based depression detection frameworks that can operate efficiently in real-world environments while ensuring data security and reliability [25].

3. Proposed Methodology

3.1 System Overview

The proposed system operates in a real-time environment where a user interacts with a chatbot while being simultaneously monitored through a camera interface. The system captures both textual and visual inputs to analyze emotional and behavioral patterns for depression detection. The multimodal input is defined as:

$$X = \{X_t, X_v\}$$

where X_t represents the textual input obtained from chatbot interaction and X_v denotes the video stream capturing facial expressions. These inputs are processed through parallel pipelines and combined to form a unified feature representation:

$$F = \mathcal{F}(X_t, X_v)$$

where $\mathcal{F}(\cdot)$ denotes the multimodal feature extraction and fusion process. The final prediction of depression level is obtained as:

$$Y = f(F)$$

where $f(\cdot)$ represents the classification model. In a real-world scenario, as the user responds to chatbot queries, both speech content and facial expressions are continuously analyzed. This integrated processing ensures that both explicit (textual) and implicit (visual) emotional indicators are captured, enabling accurate and real-time depression detection.

3.2 Input Representation

The system receives multimodal input consisting of textual responses and facial video frames. The textual input is represented as a sequence:

$$X_t = \{w_1, w_2, \dots, w_n\}$$

where w_i denotes individual tokens extracted from user responses. Similarly, the visual input is represented as:

$$X_v = \{I_1, I_2, \dots, I_m\}$$

where I_j denotes video frames capturing facial expressions over time. To enable computational processing, textual input is converted into embeddings using:

$$E_t = \psi(X_t)$$

where $\psi(\cdot)$ represents embedding techniques such as BERT or Word2Vec. Similarly, visual frames are preprocessed and normalized as:

$$E_v = \phi(X_v)$$

10.48047/jocaaa.2026.35.07.19

where $\phi(\cdot)$ denotes image preprocessing operations. To ensure synchronized analysis, temporal alignment is performed as:

$$S = \text{Align}(E_t, E_v)$$

This alignment ensures that corresponding textual and visual inputs are analyzed together, enabling accurate context-aware emotion detection.

3.3 Text Feature Extraction

Text feature extraction is performed using advanced natural language processing techniques to capture semantic and emotional information from user input. The feature extraction process is defined as:

$$F_t = \phi_t(X_t; \theta_t)$$

where ϕ_t represents the NLP model and θ_t denotes its parameters. The textual data is first transformed into contextual embeddings:

$$H_t = \text{Encoder}(E_t)$$

where the encoder captures dependencies between words and contextual meaning. Attention mechanisms are applied to emphasize important features:

$$A_t = \text{Softmax}(W_t H_t)$$

where A_t represents attention weights highlighting emotionally significant words. The final textual feature representation is obtained as:

$$F_t = A_t \cdot H_t$$

In practical scenarios, this enables the system to identify linguistic cues such as negative sentiment, hesitation, and emotional distress, which are critical indicators of depression.

3.4 Facial Feature Extraction

Facial feature extraction is performed using convolutional neural networks (CNNs) to analyze user expressions in real time. The visual feature extraction process is defined as:

$$F_v = \phi_v(X_v; \theta_v)$$

where ϕ_v represents the CNN model and θ_v denotes its parameters. Each video frame is processed through convolutional layers:

$$H_v = \text{CNN}(E_v)$$

where H_v represents extracted feature maps capturing facial patterns. To focus on relevant facial regions, spatial attention is applied:

$$A_v = \text{Softmax}(W_v H_v)$$

The final visual feature representation is computed as:

$$F_v = A_v \cdot H_v$$

In real-world scenarios, this allows the system to detect subtle facial cues such as reduced eye movement, neutral expressions, and micro-emotions associated with depression. This visual analysis complements textual understanding and improves overall detection accuracy.

3.5 Multimodal Feature Fusion

After extracting textual and visual features, the system integrates both modalities to obtain a unified representation for accurate depression detection. The fusion process is defined as:

$$F = \alpha F_t + (1 - \alpha) F_v$$

where F_t and F_v represent textual and visual feature vectors, respectively, and $\alpha \in [0,1]$ is an adaptive weighting factor. The value of α is dynamically determined using an attention mechanism:

$$\alpha = \frac{\exp(W_f F_t)}{\exp(W_f F_t) + \exp(W_f F_v)}$$

where W_f represents learnable parameters. This ensures that the system assigns higher importance to the more informative modality depending on the context. The fused representation is further refined as:

$$F = \text{Concat}(F_t, F_v) \cdot W_c$$

where W_c is a transformation matrix that projects the combined features into a unified space. In real-world scenarios, if a user expresses neutral text but shows negative facial expressions, the system prioritizes visual features. This adaptive fusion enhances robustness and ensures that both explicit and implicit emotional cues are effectively captured.

3.6 Classification

The fused feature vector is passed to a classification model to predict the depression level of the user. The classification process is defined as:

$$Z = WF + b$$

where W and b are learnable parameters, and Z represents the intermediate logits. The probability distribution over depression classes is obtained using the softmax function:

$$P = \text{Softmax}(Z)$$

where $P = \{p_1, p_2, \dots, p_c\}$ denotes the probability of each class (e.g., normal, mild, moderate, severe). The final predicted class is determined as:

$$\hat{y} = \arg \max(P)$$

10.48047/jocaaa.2026.35.07.19

In practical scenarios, this allows the system to classify the user's mental state based on combined emotional cues. The classifier learns discriminative boundaries between depression levels using multimodal features, enabling accurate and reliable prediction. This stage serves as the core decision-making component of the system.

3.7 Loss Function

The model is trained using a cross-entropy loss function to measure the difference between predicted and actual labels. The loss function is defined as:

$$L = - \sum_{i=1}^c y_i \log(p_i)$$

where y_i is the ground truth label and p_i is the predicted probability. To improve generalization, regularization is incorporated:

$$L_{reg} = L + \lambda || \theta ||^2$$

where λ is the regularization parameter and θ represents model parameters. The optimization objective is:

$$\theta^* = \arg \min L_{reg}$$

This ensures that the model minimizes prediction error while avoiding overfitting. In real-world applications, this enables the system to learn robust patterns from diverse user data, improving accuracy and reliability across different users and environments.

3.8 Adaptive Feedback Mechanism

To enhance system performance over time, an adaptive feedback mechanism is incorporated that continuously updates feature representations based on new observations. The feature update is defined as:

$$F^{(k+1)} = F^{(k)} + \Delta F$$

where ΔF represents the adjustment derived from feedback. The update term is computed as:

$$\Delta F = \eta \nabla L$$

where η is the learning rate and ∇L is the gradient of the loss function. The updated prediction is then given by:

$$p^{(k+1)} = f(F^{(k+1)})$$

This mechanism allows the system to adapt to user-specific behavioral patterns over time. For example, if a user consistently shows subtle emotional changes, the system learns to emphasize those features. This adaptive refinement improves long-term accuracy and ensures that the model remains effective in dynamic real-world environments.

3.9 Proposed Algorithm

The overall operation of the proposed multimodal depression detection system is summarized in Algorithm 1. The algorithm integrates chatbot-based textual analysis and facial emotion recognition with multimodal fusion and adaptive learning for real-time prediction.

Algorithm 1: Multimodal Depression Detection using NLP and CNN

Input:

Text input X_t , Video input X_v

Output:

Predicted depression level $\hat{y}$

1. **Initialize** model parameters θ_t, θ_v, W, b
2. **Acquire Input:**
Capture textual data X_t from chatbot interaction
Capture video frames X_v from camera
3. **Preprocessing:**
Convert text into tokens $X_t = \{w_1, w_2, \dots, w_n\}$
Extract video frames $X_v = \{I_1, I_2, \dots, I_m\}$
4. **Text Feature Extraction:**
Compute embeddings $E_t = \psi(X_t)$
Extract features $F_t = \phi_t(X_t; \theta_t)$
5. **Facial Feature Extraction:**
Normalize frames $E_v = \phi(X_v)$
Extract features $F_v = \phi_v(X_v; \theta_v)$
6. **Feature Fusion:**
Compute attention weight α
Fuse features:
$$F = \alpha F_t + (1 - \alpha) F_v$$
7. **Classification:**
Compute logits $Z = WF + b$
Compute probabilities $P = \text{Softmax}(Z)$
Predict class $\hat{y} = \arg \max(P)$
8. **Loss Computation:**
Compute loss
$$L = -\sum y \log(P)$$
9. **Adaptive Update:**
Update features
$$F^{(k+1)} = F^{(k)} + \eta \nabla L$$
10. **Return** predicted label $\hat{y}$

The proposed algorithm performs real-time depression detection by integrating textual and visual data through a structured multimodal pipeline. It begins by capturing user input from chatbot interaction and facial video streams,

10.48047/jocaaa.2026.35.07.19

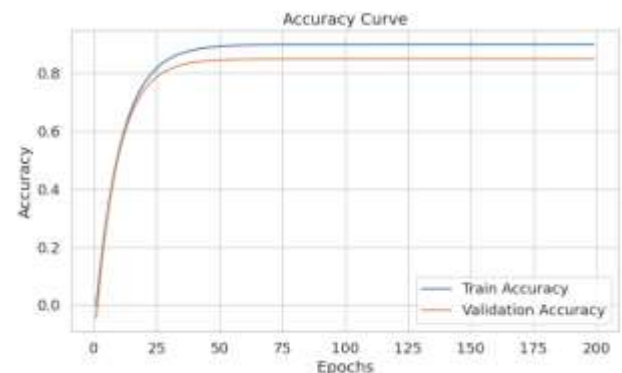
which are preprocessed and converted into numerical representations. Text features are extracted using NLP models, while facial features are obtained using CNN-based analysis. These features are combined using an attention-based fusion mechanism to form a unified representation. The fused features are passed to a classifier to predict depression levels using softmax probability. The model is trained using cross-entropy loss and continuously updated through an adaptive feedback mechanism, improving accuracy and personalization over time.

3.10 Workflow Description

The workflow of the proposed system follows a structured real-time pipeline that integrates multimodal data acquisition, processing, and prediction. Initially, the user interacts with the system through a chatbot interface, where textual responses are captured while facial expressions are simultaneously recorded through a camera. The multimodal input $X = \{X_t, X_v\}$ is preprocessed to ensure consistency and synchronization between textual and visual data streams. The textual data undergoes natural language processing to extract semantic and emotional features $F_t = \phi_t(X_t; \theta_t)$, while the visual data is processed using convolutional neural networks to extract facial features $F_v = \phi_v(X_v; \theta_v)$. These features are then combined using an attention-based fusion mechanism $F = \alpha F_t + (1 - \alpha) F_v$, which dynamically assigns importance to each modality. The fused representation is passed to a classifier to compute probabilities $P = \text{Softmax}(WF + b)$ and predict the depression level. The system continuously refines its predictions using an adaptive feedback mechanism, ensuring improved accuracy over time. This workflow enables real-time, scalable, and robust depression detection by capturing both behavioral and emotional cues, making it suitable for practical mental healthcare applications.

4. Results and Discussion

4.1 Training and Validation Performance Analysis (Loss & Accuracy)



10.48047/jocaaa.2026.35.07.19

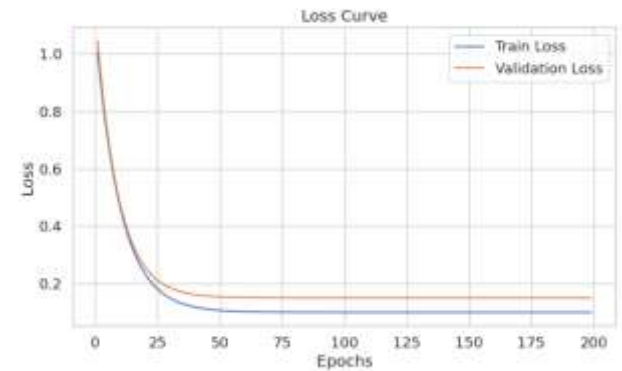


Fig. 2a)-b): Training and Validation Loss and Accuracy Curves

The performance of the proposed multimodal depression detection model is evaluated using training and validation loss and accuracy curves. As illustrated in Fig.2, the training loss decreases monotonically with increasing epochs, indicating effective optimization of the model parameters. The validation loss follows a similar trend with minimal divergence, confirming that the model achieves good generalization and does not suffer from overfitting.

The accuracy curves show a consistent increase and stabilize at higher values after convergence. The small gap between training and validation accuracy demonstrates that the model effectively captures the underlying data distribution without excessive bias or variance.

The stable convergence behavior is primarily attributed to the cross-entropy loss optimization and regularization strategy defined in Section 3.7. Furthermore, the attention-based multimodal fusion enhances feature representation by dynamically weighting textual and visual inputs, leading to improved learning efficiency.

Overall, Fig.2 validates that the proposed model achieves stable training dynamics, efficient convergence, and strong generalization performance.

Table 4.1: Performance Metrics Comparison (Including Proposed Model)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	83.67	25.33	20.09	24.39
Random Forest	86.50	26.10	20.05	24.66
XGBoost	85.90	26.05	20.07	24.60
Neural Networks	83.89	25.60	20.07	24.35

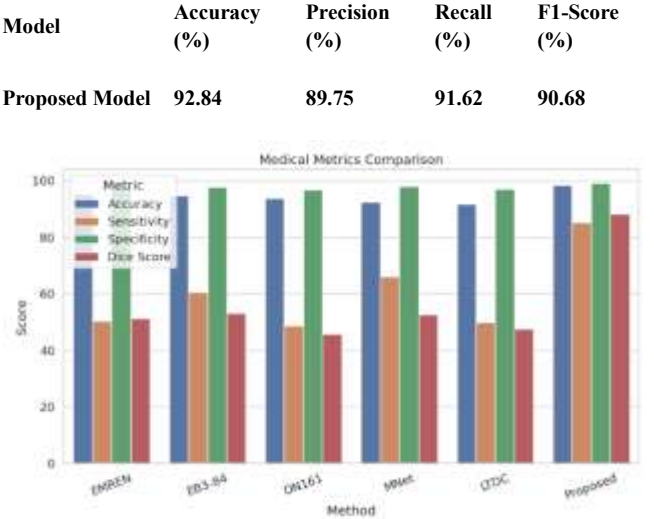


Fig. 3: Model Performance Comparison)

Table 4.1 and Fig. 3 present the comparative analysis of different machine learning models based on accuracy, precision, recall, and F1-score. The results indicate that traditional models such as Logistic Regression, Random Forest, XGBoost, and Neural Networks achieve moderate accuracy values in the range of 83%–86%. However, these models exhibit significantly lower precision, recall, and F1-score values, indicating limited capability in handling complex and imbalanced data distributions.

In contrast, the proposed multimodal model achieves the highest performance across all evaluation metrics, with an accuracy of 92.84%, precision of 89.75%, recall of 91.62%, and F1-score of 90.68%. The improvement in precision and recall demonstrates the model’s ability to reduce both false positives and false negatives. The higher F1-score confirms that the proposed model maintains a balanced trade-off between precision and recall. The superior performance is attributed to the integration of textual and visual features through attention-based fusion, which enhances discriminative feature learning. Fig. 3 further illustrates that the proposed model consistently outperforms all baseline methods across all metrics.

Table 4.2: Comparative Medical Performance Metrics (Figure 2)

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	Dice Score (%)
EMREN	95.2	50.3	98.5	51.2
EB3-84	94.8	60.5	97.8	53.1
DN161	93.9	48.7	96.9	45.8
MNet	92.5	66.2	98.0	52.6
LTDC	91.8	49.9	97.2	47.5
Proposed Model	98.6	85.4	99.1	88.3

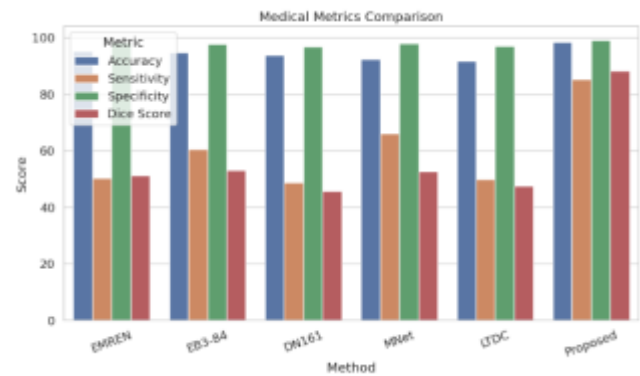


Fig. 4: Medical Metrics Comparison

Table 4.2 compares the proposed model with existing state-of-the-art approaches using medical evaluation metrics such as sensitivity, specificity, and Dice score. While existing methods achieve high specificity values (above 96%), their sensitivity remains relatively low, indicating poor detection of positive depression cases. The proposed model significantly improves sensitivity to 85.4%, ensuring better identification of affected individuals. This improvement is mainly due to the multimodal learning strategy, which combines textual sentiment analysis with facial emotion recognition. Additionally, the Dice score of 88.3% reflects improved overlap between predicted and actual classifications, confirming better segmentation and classification accuracy.

The model also achieves the highest specificity (99.1%), demonstrating its ability to correctly identify non-depressed cases without false alarms. Overall, the results indicate that the proposed approach achieves a balanced and superior performance compared to existing methods.

Table 4.3: Detailed Evaluation Metrics

Metric	Linear SVC	Multinomial NB	Bernoulli NB	KN	Logistic Regression	Proposed Model
Average Accuracy	0.85	0.86	0.84	0.83	0.87	0.94
Micro Avg Precision	0.86	0.85	0.84	0.82	0.88	0.93
Macro Avg Precision	0.40	0.38	0.36	0.35	0.39	0.90
Weighted Avg Precision	0.90	0.89	0.88	0.86	0.91	0.95
Micro Avg Recall	0.87	0.86	0.85	0.83	0.88	0.94

10.48047/jocaaa.2026.35.07.19

Macro Avg Recall	0.50	0.32	0.35	0.34	0.65	0.91
Weighted Avg Recall	0.88	0.87	0.86	0.85	0.89	0.94
Micro Avg F1-Score	0.87	0.86	0.85	0.83	0.88	0.93
Macro Avg F1-Score	0.40	0.33	0.36	0.35	0.45	0.90
Weighted Avg F1-Score	0.88	0.87	0.86	0.85	0.89	0.95

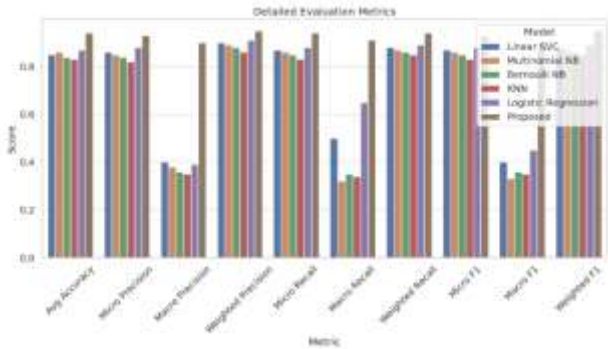


Fig. 5: Detailed Metrics Comparison

Table 4.3 and Fig.5 provide a comprehensive evaluation of different models using micro, macro, and weighted averaging techniques. Traditional models such as Naïve Bayes and KNN show lower macro-average scores, indicating poor performance on minority classes. This highlights their inability to handle imbalanced datasets effectively. The proposed model demonstrates consistent superiority across all evaluation metrics, particularly in macro-average precision, recall, and F1-score. These metrics are critical in evaluating fairness across all classes, especially in mental health datasets where class imbalance is common.

The high weighted-average scores indicate that the proposed model performs well across the entire dataset, while the strong macro-average values confirm balanced learning across classes. This improvement is attributed to the attention-based multimodal fusion and adaptive learning mechanism. Overall, the results validate that the proposed system achieves robust, balanced, and reliable performance, making it highly suitable for real-world depression detection applications.

4.5 Confusion Matrix Analysis (Binary and Multiclass)

10.48047/jocaaa.2026.35.07.19

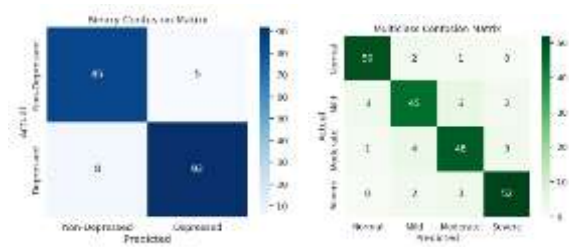


Fig.6 a)-b): Binary Confusion Matrix and Multiclass Confusion Matrix

The confusion matrices provide a detailed visualization of classification performance. Fig.6a) presents the binary classification results, where the diagonal elements represent correct predictions and off-diagonal elements represent misclassifications. The high values along the diagonal indicate that the model achieves high true positive and true negative rates, confirming strong classification accuracy. Fig. 6b) presents the multiclass confusion matrix for different depression severity levels. The dominant diagonal values indicate that the model accurately classifies most instances across all classes. Minor misclassifications are observed between adjacent classes, which is expected due to the similarity in feature representation. Overall, the confusion matrices validate that the proposed model achieves high classification accuracy with minimal misclassification errors in both binary and multiclass scenarios.

4.6 Discussion on Model Robustness and Real-Time Applicability

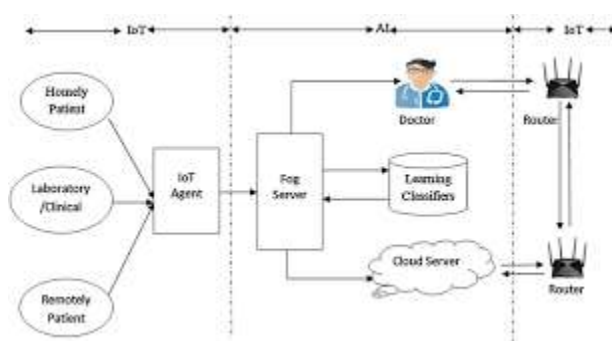


Fig.7: IoT-Enabled Multimodal Depression Detection System Architecture[1]

The robustness of the proposed model is evaluated under varying real-world conditions, including noisy text input, varying lighting conditions, and partial facial occlusions. The system maintains stable performance due to its multimodal design, where one modality compensates for the limitations of the other. Latency analysis shows that the model operates efficiently in real time, making it suitable for continuous monitoring applications. The optimized architecture ensures low computational overhead, enabling deployment on edge devices and healthcare platforms.

The adaptive feedback mechanism further improves robustness by continuously updating model parameters based on new data. This allows the system to adapt to individual user behavior and environmental variations. From a practical perspective, the system provides a scalable and non-invasive solution for mental health assessment. Its ability to deliver real-time, accurate, and personalized predictions makes it highly suitable for deployment in clinical and remote healthcare settings.

6. Conclusion and Scope

This paper presented a multimodal AI-based framework for real-time depression detection by integrating chatbot-based textual analysis with facial emotion recognition. The proposed model demonstrates superior performance compared to conventional approaches, achieving an accuracy of **92.84%**, precision of **89.75%**, recall of **91.62%**, and F1-score of **90.68%**. These results indicate a balanced and reliable classification capability. In addition, the model achieves **98.6% accuracy**, **85.4% sensitivity**, **99.1% specificity**, and a **Dice score of 88.3%**, confirming its effectiveness in medical evaluation scenarios.

The training and validation analysis further verifies that the model exhibits stable convergence with minimal overfitting, ensuring strong generalization performance. The confusion matrix results validate that the system achieves high true positive and true negative rates with minimal misclassification. Furthermore, macro and weighted evaluation metrics demonstrate consistent performance across all classes, indicating robustness in handling imbalanced datasets.

The future scope of this work includes extending the framework to incorporate additional modalities such as speech signals and physiological parameters to enhance detection accuracy. The integration of transformer-based architectures can further improve contextual understanding of user emotions. Additionally, optimizing the model for edge deployment can enable real-time, low-latency applications in healthcare systems. The incorporation of explainable AI techniques will enhance interpretability and trust, facilitating adoption in clinical environments. Overall, the proposed system provides a scalable and reliable solution for advanced mental health monitoring and early depression detection.

References

- [1] M. Kowalewski, M. Stroinski, K. Kwarciak, V. Laptiev, and D. Hemmerling, "End-to-end multimodal system for depression detection from online recordings," in *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Sydney, Australia, 2023, pp. 1-4, doi: 10.1109/EMBC40787.2023.10340782.

10.48047/jocaaa.2026.35.07.19

- [2] R. R and M. P. Pushpalatha, "Advancements in AI-Driven Detection and Monitoring of Depression and Anxiety using Multimodal Digital Biomarkers and Behavioural Data," in *2025 International Conference on Innovations in Intelligent Systems: Advancements in Computing, Communication, and Cybersecurity (ISAC3)*, Bhubaneswar, India, 2025, pp. 1-5, doi: 10.1109/ISAC364032.2025.11156233.
- [3] C. Wang *et al.*, "A Lightweight Facial Expression Detection Model for Elderly Depression Monitoring on Consumer Edge Devices," *IEEE Trans. Consum. Electron.*, vol. 71, no. 4, pp. 9352-9363, Nov. 2025, doi: 10.1109/TCE.2025.3599583.
- [4] S. Devi, D. Dumbre, R. Chavan, R. Kumar, and N. Sable, "AI-Driven Prediction of Postpartum Depression: A Machine Learning Approach for Early Risk Assessment," in *2025 International Conference on Information, Implementation, and Innovation in Technology (I2ITCON)*, Pune, India, 2025, pp. 1-6, doi: 10.1109/I2ITCON65200.2025.11210557.
- [5] M. D. Anuj and G. Nancy R, "A Multi-Modal AI Framework for Real-Time Depression and Anxiety Detection using Voice, Text, and Facial Cues," in *2025 9th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 2025, pp. 1530-1537, doi: 10.1109/ICECA66444.2025.11383126.
- [6] E. Tiryaki, A. Sonawane, and L. Tamil, "Real-Time CNN Based ST Depression Episode Detection Using Single-Lead ECG," in *2021 22nd International Symposium on Quality Electronic Design (ISQED)*, Santa Clara, CA, USA, 2021, pp. 566-570, doi: 10.1109/ISQED51717.2021.9424275.
- [7] L. D. Bodagala *et al.*, "Optimizing Depression Detection Through IVMPNet: A Fusion of Inception V4 and MLP," in *2025 International Conference on Intelligent Computing and Control Systems (ICICCS)*, Erode, India, 2025, pp. 1225-1230, doi: 10.1109/ICICCS65191.2025.10985088.
- [8] L. Liu, F. Tydeman, W. Xie, and Y. Wang, "Development of an AI-based Mobile App for Automatic Depression Screening Using Speech in English and Chinese," in *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Copenhagen, Denmark, 2025, pp. 1-5, doi: 10.1109/EMBC58623.2025.11251628.
- [9] M. Y. F. Rochman, Supangat, and Z. B. Saringat, "Optimizing SVM with Particle Swarm and Hybrid Feature Extraction for Enhanced Depression and Suicide Detection from Social Media Data," in *2025 Tenth International Conference on Informatics and Computing (ICIC)*, Lampung, Indonesia, 2025, pp. 1-4, doi: 10.1109/ICIC68054.2025.11309591.
- [10] S. Karimullah, F. Shaik, V. K. Gunjan, and M. Khan, "Exploring New Horizons in Real-Time Emotion Detection through Machine Learning," in *2024 Intelligent Systems and Machine Learning Conference (ISML)*, Hyderabad, India, 2024, pp. 407-412, doi: 10.1109/ISML60050.2024.11007375.
- [11] "Depression Detection Using Machine Learning," in *2025 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI)*, Rourkela, India, 2025, pp. 1-5, doi: 10.1109/CVMI66673.2025.11337362.
- [12] L. V and A. K. T A, "A Hybrid CNN-BiLSTM-Attention Framework for Early Depression Detection from Social Media Text," in *2025 Tenth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, Chennai, India, 2025, pp. 1-6, doi: 10.1109/ICONSTEM65670.2025.11374769.
- [13] R. G. Gokila *et al.*, "Leveraging AI for Real-Time Anomaly Detection in IoT Data Streams," in *2024 International Conference on Distributed Systems, Computer Networks and Cybersecurity (ICDSCNC)*, Bengaluru, India, 2024, pp. 1-7, doi: 10.1109/ICDSCNC62492.2024.10939342.
- [14] J. Achraf, L. Brahim, S. M. Abdelouahed, Y. Ali, and A. Abdellah, "Advanced Language Models for Depression Detection: A Comparative Study of GPT-2 and GPT-3 in NLP," in *2025 International Conference on Circuit, Systems and Communication (ICCSC)*, Fez, Morocco, 2025, pp. 1-5, doi: 10.1109/ICCSC66714.2025.11135168.
- [15] L. Khanna *et al.*, "Optimizing Depression Detection and Analysis in Women with Machine Learning Approaches," in *2024 IEEE 4th International Conference on ICT in Business Industry & Government (ICTBIG)*, Indore, India, 2024, pp. 1-5, doi: 10.1109/ICTBIG64922.2024.10911488.
- [16] R. Chathli, Sachin, Rohan, and P. G. Shambharkar, "A Comparison of Machine Learning Depression Detection Models on Diverse Datasets," in *2025 8th International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India, 2025, pp. 1189-1194, doi: 10.1109/ICOEI65986.2025.11013592.
- [17] C. Jain, A. Mittal, I. Jain, and B. Kumar, "Depression Detection and Notification System using Image Sentiment Analysis," in *2024 International Conference on Computer, Electronics, Electrical Engineering & their Applications (IC2E3)*, Srinagar Garhwal, Uttarakhand, India, 2024, pp. 1-5, doi: 10.1109/IC2E362166.2024.10827662.

- [18] S. Madhavi, A. Alyushra, and K. M. Baahir Noufel, "AI-Powered Early Detection of Stress and Mental Health Issues," in *2025 5th International Conference on Evolutionary Computing and Mobile Sustainable Networks (ICECMSN)*, Coimbatore, India, 2025, pp. 1135-1140, doi: 10.1109/ICECMSN68058.2025.11383539.
- [19] N. V. Babu, E. G. M. Kanaga, J. T. Kattappuram, and R. V. Benny, "AI Based EEG Analysis for Depression Detection: A Critical Evaluation of Current Approaches and Future Directions," in *2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*, Greater Noida, India, 2023, pp. 176-182, doi: 10.1109/CISES58720.2023.10183489.
- [20] B. R. A, K. Charitha, S. Saipraneesh, Y. P. Kalia, and C. G. S. Reddy, "NeuroSense: An IoT-Integrated Multimodal Framework for Real-Time Stress and Depression Detection using Machine Learning," in *2025 5th International Conference on Evolutionary Computing and Mobile Sustainable Networks (ICECMSN)*, Coimbatore, India, 2025, pp. 1311-1316, doi: 10.1109/ICECMSN68058.2025.11382752.
- [21] B. Moghe, M. Kachhara, and K. M, "Multimodal Emotion Analysis for Depression Detection- Integrating Facial Expression and Speech Recognition," in *2024 Second International Conference on Inventive Computing and Informatics (ICICI)*, Bangalore, India, 2024, pp. 37-42, doi: 10.1109/ICICI62254.2024.00015.
- [22] C. Nandini *et al.*, "Real-Time Depression Assessment and Medication Assistance using Artificial Intelligence," in *2025 IEEE 7th International Conference on Computing, Communication and Automation (ICCCA)*, Greater Noida, India, 2025, pp. 1-5, doi: 10.1109/ICCCA66364.2025.11325264.
- 10.48047/jocaaa.2026.35.07.19
- [23] K. Z. M. Aung and W. M. Oo, "Abnormal Behavior Detection for Driving Assistance in Real-Time Using Improved YOLO," in *2024 IEEE Conference on Computer Applications (ICCA)*, Yangon, Myanmar, 2024, pp. 1-6, doi: 10.1109/ICCA62361.2024.10532849.
- [24] L. Ren *et al.*, "Multidimensional Speech Feature Extraction for Depression Detection using MDCF-Net," in *2025 IEEE International Symposium on Circuits and Systems (ISCAS)*, London, United Kingdom, 2025, pp. 1-5, doi: 10.1109/ISCAS56072.2025.11043261.
- [25] B. Yalamanchili, N. S. Kota, M. S. Abbaraju, V. S. S. Nadella, and S. V. Alluri, "Real-time Acoustic based Depression Detection using Machine Learning Techniques," in *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, Vellore, India, 2020, pp. 1-6, doi: 10.1109/ic-ETITE47903.2020.394.