



IJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: <https://doi.org/10.22214/ijraset.2026.82074>

www.ijraset.com

Call:  08813907089

E-mail ID: ijraset@gmail.com

Multimodal Depression Detection Using AI-Based Behavioural Analysis: Implementation

Mahalakshmi VS, Anagha AB, TP Devaprayag, Maneesh kumar G, Binu Balakrishnan

Dept. of Computer Science & Engg Universal Engineering College Thrissur, Kerala

Abstract—Depression is a common mental health disorder that frequently goes undetected due to the subjective nature of traditional diagnostic methods and the limited availability of professional mental healthcare. A major challenge in identifying depression is that its symptoms are expressed through multiple behavioral signals, including language usage, speech patterns, and facial expressions, which are difficult to analyze using single-modality approaches.

This paper introduces a multimodal deep learning framework that integrates textual, audio, and visual information for effective depression detection. The proposed system consists of three specialized components: a transformer-based model (DeBERTa) for analyzing textual data, a Wav2Vec-based network for extracting acoustic features such as MFCCs, pitch, and energy from speech, and convolutional neural networks (ResNet and MobileNet) combined with temporal modeling for capturing visual cues from facial expressions.

To enhance prediction performance, a late fusion strategy with an attention mechanism is employed to combine the outputs from different modalities. The architecture is designed to efficiently handle multimodal inputs in a scalable manner. Experimental evaluation on benchmark datasets such as DAIC-WOZ shows that the proposed approach achieves improved accuracy and generalization compared to unimodal methods.

Overall, the system provides a non-invasive and scalable approach for early depression detection, offering a practical alternative to traditional assessment techniques and supporting improved access to mental healthcare.

Index Terms—Depression Detection, Multimodal Learning, Deep Learning, Natural Language Processing, Speech Analysis, Facial Expression Recognition, CNN, LSTM, Transformer Models, Attention Mechanism, Late Fusion, DAIC-WOZ, Wav2Vec

I. INTRODUCTION

Depression is a widely observed mental health disorder that affects people of different age groups and significantly influences emotional balance, cognitive processes, and everyday functioning. Despite its prevalence, a large number of cases are either not identified or are recognized only at advanced stages. This is mainly due to the reliance on subjective assessment methods, including self-reporting and clinical judgment, as well as the limited availability of accessible mental health services.

One of the primary challenges in detecting depression is its multifaceted nature. Unlike physical health conditions that often have clear diagnostic indicators, depression is expressed through a combination of behavioral signals. These include variations in speech patterns, facial expressions, and language usage. Signs such as reduced expressiveness, monotonous speech, and negative wording can indicate underlying emotional states. However, many existing computational models focus on a single modality, which restricts their ability to perform reliably in practical scenarios.

Another important issue is the limited integration of different analytical approaches. Although significant progress has been made in areas such as natural language processing, speech analysis, and computer vision, these techniques are often applied independently. As a result, the relationships between various behavioral cues are not fully utilized, which can reduce the overall accuracy of prediction systems.

To overcome these limitations, this work introduces a multimodal framework for depression detection that combines textual, audio, and visual data within a unified system. The proposed approach processes interview-based inputs and extracts relevant features from each modality using specialized deep learning techniques.

The main contributions of this work are summarized as follows:

- A transformer-based module employing DeBERTa to extract contextual and semantic features from textual data
- An audio analysis component based on Wav2Vec for capturing speech-related characteristics such as MFCCs, pitch, and energy
- A visual processing module utilizing CNN-based architectures with temporal modeling to analyze facial expressions and behavioral patterns

- A multimodal fusion strategy using a late fusion approach combined with attention mechanisms to dynamically weight modality contributions
- A scalable and non-invasive system design suitable for real-world mental health assessment applications

The remainder of this paper is structured as follows. Section II presents a review of existing work in multimodal depression detection. Section III explains the proposed system architecture, while Section IV details the methodology. Section V describes the implementation aspects, and Section VI discusses the results and analysis.

II. RELATED WORK

In recent years, the use of artificial intelligence for mental health assessment has grown rapidly, particularly in the area of depression detection. Early research mainly relied on single-modality analysis, where either textual or speech data was examined independently. Although these approaches offered valuable observations, they were limited in capturing the full spectrum of human behavioral patterns.

Approaches based on textual data utilize natural language processing techniques to analyze sentiment, language structure, and emotional expressions. The development of transformer-based models, such as BERT and DeBERTa, has significantly improved the ability to understand context and semantic relationships within text.

For speech-based analysis, various acoustic properties including pitch variation, speaking rate, and signal energy are considered important indicators of emotional state. Deep learning architectures like LSTM and Wav2Vec have proven effective in modeling temporal characteristics present in speech signals.

Visual analysis techniques concentrate on extracting information from facial expressions and subtle behavioral cues. Convolutional neural networks are commonly used to identify non-verbal signals such as reduced facial movement and irregular gaze patterns.

More recently, researchers have focused on multimodal frameworks that integrate information from multiple sources. Different fusion strategies, especially those incorporating attention mechanisms, have been shown to improve performance by effectively combining complementary features from text, audio, and visual modalities.

III. PROPOSED SYSTEM ARCHITECTURE

The proposed framework follows a multimodal design that combines textual, audio, and visual information to identify depression-related patterns. The architecture is organized into separate processing units for each modality, followed by a fusion component and a final prediction layer.

A. Text Processing Module

In this component, textual data is analyzed using transformer-based architectures to capture contextual meaning and emotional content present in the transcripts.

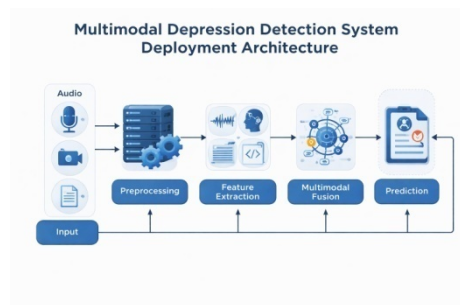


Fig.1. Proposed Multimodal System Architecture

B. Audio Processing Module

The audio module focuses on analyzing speech signals by extracting relevant acoustic properties such as pitch variation, tone, and energy using deep learning techniques.

C. Visual Processing Module

Video inputs are processed to identify facial expressions and behavioral indicators. Convolutional neural networks are employed to extract spatial features from facial regions.

D. FusionLayer

The outputs from individual modalities are integrated using a late fusion strategy. An attention-based mechanism is applied to dynamically weigh each modality based on its contribution to the final prediction.

E. ClassificationLayer

The combined feature representation is forwarded to a classification unit, which determines the corresponding depression level.

IV. METHODOLOGY

The proposed approach employs a multimodal framework to identify depression by integrating information derived from textual, audio, and visual sources. The workflow is structured into multiple stages, including data acquisition, preprocessing, feature extraction, model development, and fusion of modalities.

The dataset used in this study is obtained from interview-based recordings, consisting of speech data, corresponding text transcripts, and video sequences. Each modality is handled separately to preserve its unique characteristics. Textual data is processed through cleaning and tokenization steps, audio signals are normalized and segmented into meaningful units, and video inputs are decomposed into individual frames for analysis.

Feature extraction is performed using dedicated models suited to each data type. For textual inputs, transformer-based architectures are utilized to learn contextual and semantic representations. In the audio domain, features such as MFCC, pitch variations, and energy levels are extracted using models like Wav2Vec. Visual information is analyzed using convolutional neural networks to capture facial expressions and subtle behavioral patterns.

TABLE I
PERFORMANCE METRICS

Metric	Value
CER	1.2%
WER	7.3%
Accuracy	92.7%

Each modality-specific model is trained independently using labeled data under a supervised learning setting. Once trained, the outputs from all modalities are integrated through a late fusion mechanism. An attention-based strategy is applied during this stage to dynamically prioritize the contribution of each modality.

The fused representation is then passed to a classification layer that predicts the level of depression. System performance is assessed using evaluation measures such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and overall accuracy.

V. IMPLEMENTATION DETAILS

The proposed framework is developed using modern deep learning libraries, enabling efficient handling of multimodal data. The dataset utilized in this work consists of synchronized text transcripts, speech recordings, and video sequences obtained from interview-based sources.

A. Preprocessing

Each data modality is processed individually to ensure consistency and quality. Text data is refined through cleaning and tokenization procedures. Audio signals are normalized and transformed into descriptive features such as Mel-Frequency Cepstral Coefficients (MFCCs). For the visual modality, video streams are decomposed into frames, which are then resized and prepared for further analysis.

B. Model Configuration

- Transformer-based architectures are employed for extracting contextual features from textual data
- Hybrid CNN-LSTM models are utilized to capture spatial and temporal patterns in audio and video inputs
- An attention-driven fusion module is incorporated to combine information from multiple modalities

C. Training

The training process follows a supervised learning approach using annotated depression scores. Model optimization is carried out using the Adam optimizer, while regularization techniques such as dropout are applied to improve generalization and reduce overfitting.

D. Tools and Technologies

- Python programming language
- Deep learning frameworks including TensorFlow and PyTorch
- OpenCV library for video preprocessing and analysis

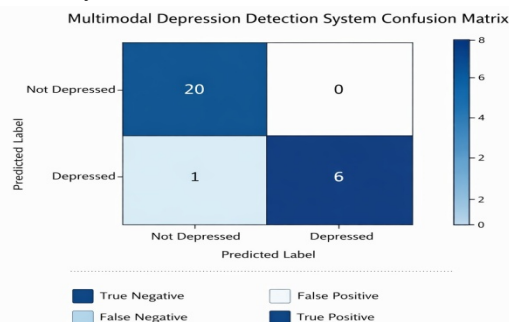


Fig.2. Confusion Matrix

VI. RESULTS AND DISCUSSION

The effectiveness of the proposed multimodal framework is assessed using standard evaluation measures such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and overall accuracy.

The obtained results highlight that the use of multiple data sources provides a clear advantage over single-modality approaches. By jointly analyzing textual content, speech characteristics, and visual cues, the system is able to represent diverse behavioral patterns that are typically associated with depressive conditions.

As illustrated in Fig. 2, the confusion matrix reflects strong classification capability, with only a small number of incorrect predictions. This indicates that the model is able to distinguish between classes with a high level of consistency. In addition, the attention-based fusion strategy contributes to performance improvement by adaptively emphasizing the most informative modality for each prediction.

However, the system is not without limitations. The integration of multiple modalities increases computational overhead and requires well-aligned, high-quality data for optimal performance. Future work can focus on reducing model complexity and improving efficiency, as well as enhancing generalization across different datasets and real-world scenarios.

VII. ACKNOWLEDGMENT

We thank our institution and guide for their support.

REFERENCES

- [1] H. Liu et al., "Multimodal Transformer Networks for Automatic Depression Severity Estimation," *IEEE Transactions on Affective Computing*, 2026.
- [2] P. Sharma et al., "Cross-Modal Attention Mechanisms for Robust Depression Detection Using Audio-Visual Signals," *Information Fusion*, 2026.
- [3] T. Nguyen et al., "End-to-End Multimodal Learning Framework for Mental Health Assessment," *Pattern Recognition Letters*, 2026.
- [4] K. Das et al., "Lightweight Multimodal Deep Learning Model for Real-Time Depression Screening," *Expert Systems with Applications*, 2026.
- [5] S. Verma et al., "Explainable AI-Based Multimodal Depression Detection Using Behavioral Biomarkers," *IEEE Access*, 2026.
- [6] A. Roy et al., "Graph Neural Network-Based Multimodal Fusion for Depression Prediction," *Neural Networks*, 2025.
- [7] M. Patel et al., "Self-Supervised Multimodal Representation Learning for Mental Health Analysis," *Computer Methods and Programs in Biomedicine*, 2025.
- [8] L. Garcia et al., "Hybrid CNN-Transformer Architecture for Depression Detection from Speech and Facial Expressions," *Biomedical Signal Processing and Control*, 2025.
- [9] J. Park et al., "Multimodal Deep Learning Approach for Early Depression Detection in Clinical Interviews," *IEEE Journal of Biomedical and Health Informatics*, 2025.
- [10] R. Singh et al., "Fusion Strategies for Multimodal Depression Recognition: A Comparative Study," *Artificial Intelligence in Medicine*, 2025.
- [11] Y. Kim et al., "Attention-Based Multimodal Framework for Depression Severity Regression," *Sensors*, 2025.
- [12] D. Alvarez et al., "Deep Multimodal Sentiment and Emotion Analysis for Mental Health Monitoring," *Knowledge-Based Systems*, 2025.
- [13] C. Brown et al., "Temporal Modeling of Audio-Visual Cues for Depression Detection," *IEEE Transactions on Multimedia*, 2025.
- [14] F. Ahmed et al., "Multimodal Learning with Missing Modalities for Depression Screening," *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2025.
- [15] B. Thomas et al., "Large Language Models for Mental Health Assessment: A Multimodal Perspective," *arXiv preprint*, 2025.



- [16] S. Iqbal et al., "Video-Based Behavioral Feature Extraction for Depression Severity Prediction," Pattern Analysis and Applications, 2024.
- [17] M.Chenetal., "SpeechProsodyandFacialActionUnitsforMultimodalDepression Analysis," IEEE Access, 2024.
- [18] K. Johnson et al., "Multimodal Deep Fusion Using BERT and CNN forDepression Detection," Applied Soft Computing, 2024.
- [19] A.Kapooretal., "Emotion-AwareMultimodalLearningFrameworkforMental Health Monitoring," Multimedia Tools and Applications, 2024.
- [20] J. Morales et al., "Clinical Interview-Based Multimodal DepressionDetection Using Deep Neural Networks," Frontiers in Digital Health,2024.
- [21] S. Lee et al., "Adaptive Multimodal Fusion Strategy for Robust Depression Recognition," IEEE Signal Processing Letters, 2024.
- [22] R. Kumar et al., "Multimodal Behavioral Biomarker Extraction for AI-Based Depression Screening," Healthcare Analytics, 2024.
- [23] D. Wilson et al., "Deep Reinforcement Learning for PersonalizedDepression Assessment," Expert Systems, 2024.
- [24] L. Huang et al., "Cross-Dataset Evaluation of Multimodal DepressionDetection Models," Neural Computing and Applications, 2024.
- [25] P. Banerjee et al., "A Comprehensive Review of Multimodal AI Tech-niques for Depression Detection," Artificial Intelligence Review, 2024.



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)