iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XII **Month of publication:** December 2025

DOI: https://doi.org/10.22214/ijraset.2025.76413

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XII Dec 2025- Available at www.ijraset.com

Multimodal Depression Detection Using AI-Based Behavioural Analysis

Anagha A B[1], T P Devaprayag[2], Mahalakshmi V S[3], Maneeshkumar G[4], Binu Balakrishnan[5]
[1,2,3,4]*B.Tech Student,* [5]*Professor, CSE Department, Universal Engineering College, Thrissur, Kerala*

Abstract: ***Depression is a major global mental health disorder that often remains undiagnosed due to the subjective nature of traditional clinical assessments and limited access to mental healthcare services. Recent advancements in artificial intelligence have enabled the development of automated systems that analyze behavioral signals for early depression detection. This paper presents a multimodal depression detection framework using AI-based behavioral analysis by integrating textual, acoustic, and visual modalities. Deep learning architectures such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and transformer-based models are employed to extract discriminative features from each modality. A fusion strategy is applied to combine complementary information and improve prediction accuracy. The proposed approach is evaluated using benchmark datasets such as DAIC-WOZ and AVEC, demonstrating that multimodal fusion significantly outperforms unimodal systems. This work highlights the effectiveness of AI-driven multimodal analysis as a reliable, non-invasive tool to support early diagnosis and monitoring of depression.***

Keywords: Multimodal depression detection, Behavioral analysis, Artificial intelligence, Deep learning, CNN, LSTM, Transformer models

I INTRODUCTION

Depression is a serious and widespread mental health disorder that affects millions of individuals globally and significantly impacts quality of life, productivity, and overall well-being. According to global health reports, depression is a leading cause of disability and is associated with increased risk of suicide and chronic illness. Despite its prevalence, depression often remains underdiagnosed and undertreated due to limitations in traditional diagnostic practices, social stigma, and lack of access to mental health professionals. Conventional depression assessment methods primarily rely on clinical interviews, self-reported questionnaires such as the PHQ-8 or PHQ-9, and observational evaluations conducted by trained clinicians. Although these methods are clinically validated, they are inherently subjective, time-consuming, and dependent on patient honesty and clinician expertise. Furthermore, many individuals avoid seeking professional help due to social stigma or geographical constraints, leading to delayed diagnosis and intervention. Human emotional and psychological states are reflected through multiple observable behavioral cues across different modalities, including speech patterns, language usage, facial expressions, and body movements. Depressed individuals often exhibit linguistic indicators such as increased use of negative emotion words, self-referential language, and reduced verbal engagement. Acoustic features such as monotonic speech, slower speaking rate, and reduced energy levels are also commonly observed. Similarly, visual cues such as reduced facial expressiveness, limited eye contact, and constrained head movements serve as important non-verbal markers of depression. Recent advancements in artificial intelligence (AI), machine learning, and deep learning have enabled the development of automated systems capable of analyzing such behavioral signals objectively. However, unimodal approaches that rely on a single data source fail to capture the multifaceted nature of depression. Emotional cues may be suppressed or inconsistently expressed in one modality while being evident in another. To overcome these limitations, multimodal depression detection systems integrate information from multiple behavioral channels to provide a more comprehensive and reliable assessment.

II LITERATURE SURVEY

1) Text-based depression detection methods (linguistic and sentiment analysis): Early research on automated depression detection primarily focused on textual data obtained from clinical interviews, questionnaires, and social media platforms. These methods analyze linguistic features such as word usage frequency, sentiment polarity, self-referential language, and syntactic complexity. Traditional machine learning models, including Support Vector Machines (SVM), Naïve Bayes, and Logistic Regression, were commonly used with handcrafted features such as TF-IDF, LIWC categories, and n-grams. While these approaches are computationally lightweight and interpretable, they struggle to capture contextual meaning and long-range semantic dependencies in language. They are also sensitive to domain shifts and linguistic diversity across users.

2) Classical speech and acoustic feature-based approaches: Speech-based depression detection systems exploit acoustic biomarkers such as reduced pitch variability, slower speech rate, lower energy, and increased pause duration, which are commonly associated with depressive states. Classical approaches extract features like Mel Frequency Cepstral Coefficients (MFCCs), pitch contours, jitter, shimmer, and spectral descriptors using toolkits such as openSMILE. These features are then classified using conventional machine learning algorithms. Although acoustic features provide objective indicators of emotional state, their performance is highly affected by background noise, microphone quality, and speaker variability. Moreover, handcrafted features may fail to generalize well across datasets and recording conditions.

3) CNN-based visual analysis (facial expression and eye behavior): Vision-based depression detection methods analyze facial expressions, eye movements, gaze patterns, and head pose from images or video sequences. Early systems relied on facial landmark detection and Action Unit (AU) analysis to identify reduced expressiveness and eye contact, which are common in depressed individuals. With the advent of deep learning, Convolutional Neural Networks (CNNs) and transfer learning models such as VGG, ResNet, and MobileNet have been employed to automatically learn discriminative visual features. These methods outperform handcrafted feature-based systems under controlled conditions; however, their accuracy degrades under poor lighting, occlusions, and variations in camera angles. Additionally, per-frame CNN classification often ignores temporal behavioral dynamics.

4) Sequential and temporal modeling approaches (LSTM, BiLSTM, 3D-CNN): Depression is not a static condition but manifests through gradual and persistent behavioral changes over time. To capture temporal dependencies, several studies integrate CNN feature extractors with Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) networks. These models analyze sequential text, speech, or video data to detect long-term depressive patterns such as sustained monotonic speech or prolonged facial inactivity. Some works also employ 3D-CNN architectures to jointly model spatial and temporal information in video data. While temporal models significantly improve robustness and detection accuracy, they introduce higher computational complexity and longer inference times, limiting their deployment in real-time or resource-constrained environments.

5) Transformer-based and attention-driven models (recent high-performance methods): Recent research has adopted transformer-based architectures to improve contextual understanding in depression detection tasks. Models such as BERT, RoBERTa, and task-oriented GPT embeddings have demonstrated strong performance in textual depression analysis by capturing long-range semantic and emotional dependencies. Attention-based multimodal fusion networks dynamically weigh different modalities based on their relevance, improving generalization across subjects and conditions. Time-aware attention mechanisms further enhance performance by emphasizing temporally significant behavioral cues. Although these methods achieve state-of-the-art results on benchmark datasets, they require large training datasets and substantial computational resources, making real-time deployment challenging without hardware acceleration.

6) Multimodal fusion approaches (text, speech, and visual integration): Multimodal depression detection systems integrate textual, acoustic, and visual cues to overcome the limitations of unimodal approaches. Feature-level, decision-level, and hybrid fusion strategies have been explored in the literature. Studies using benchmark datasets such as DAIC-WOZ and AVEC demonstrate that multimodal fusion consistently outperforms single-modality systems. Late fusion techniques are particularly effective, as they allow each modality to contribute independently while reducing noise sensitivity. However, multimodal systems face challenges related to data synchronization, missing modalities, and increased system complexity.

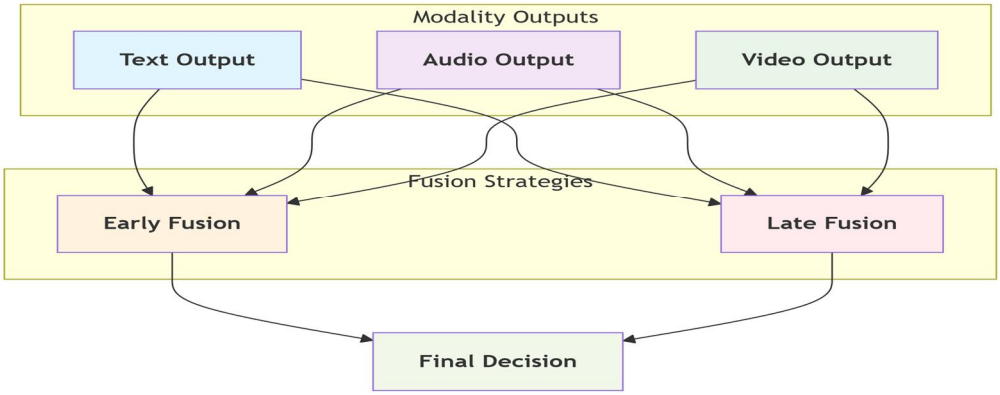


Fig7: multimodel fusion strategies

7) Social media and real-world behavioral analysis systems: Several studies analyze depression using data collected from social media platforms, including text posts, emojis, images, and user interaction patterns. Hybrid deep learning models combining CNNs and BiLSTMs have shown promising results in sentiment-based depression detection from social media data. These systems offer scalable and non-invasive monitoring but raise concerns related to data privacy, ethical usage, and potential bias due to demographic and cultural variations in online behavior.

III PROPOSED SYSTEM

The proposed multimodal depression detection system is designed to analyze behavioral indicators from text, speech, and facial expressions using AI-based techniques. The system architecture follows a modular pipeline consisting of data acquisition, modality-specific feature extraction, multimodal fusion, and depression classification.

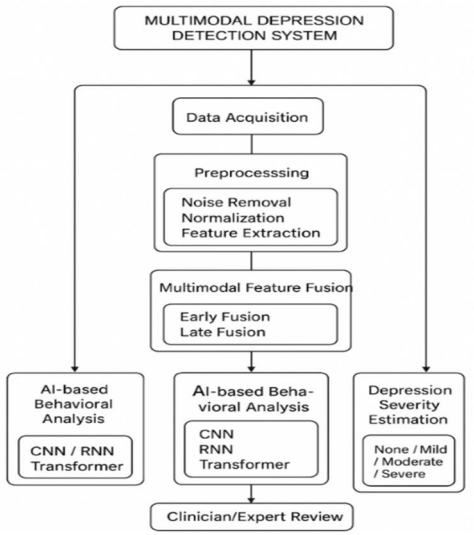


Fig.1 System Architecture

Each modality is processed independently using specialized deep learning models to ensure optimal feature representation. The extracted features are then combined using a fusion strategy to generate a final prediction regarding the individual's depressive state.

1) Textual Modality Processing: Textual data is obtained from clinical interview transcripts, questionnaires, or online social media content. Preprocessing steps include tokenization, stop-word removal, lemmatization, and normalization. Transformer-based language models such as BERT or task-oriented embeddings are employed to capture semantic meaning, emotional tone, and contextual dependencies within the text. These models provide rich representations that significantly improve depression detection performance.

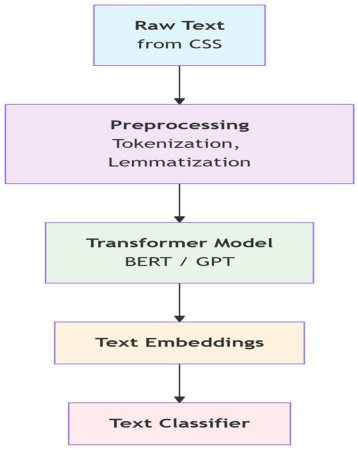


Fig4:text processing pipeline

2) Acoustic Modality Processing: Speech data is processed to extract emotion-related acoustic features including MFCCs, pitch, energy, spectral flux, and pause duration. These features are extracted using openSMILE and fed into LSTM or BiLSTM networks to model temporal speech patterns. Sequential modeling is crucial for identifying subtle and evolving depressive speech characteristics.

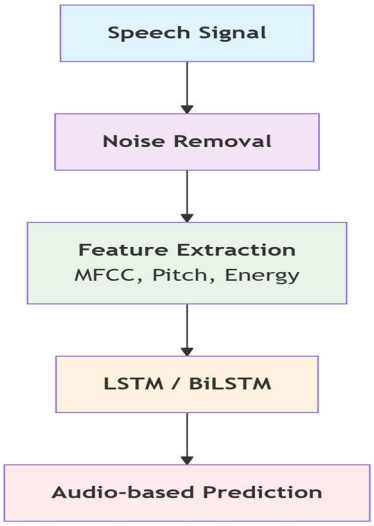


Fig5: audio processing pipeline

3) Visual Modality Processing: Visual analysis involves processing facial video frames to extract facial landmarks, gaze direction, head pose, and Action Units. CNN-based architectures such as ResNet or MobileNet are used to extract spatial features, while temporal dependencies are modeled using recurrent or attention-based layers. This approach captures both static facial expressions and dynamic behavioral changes over time.

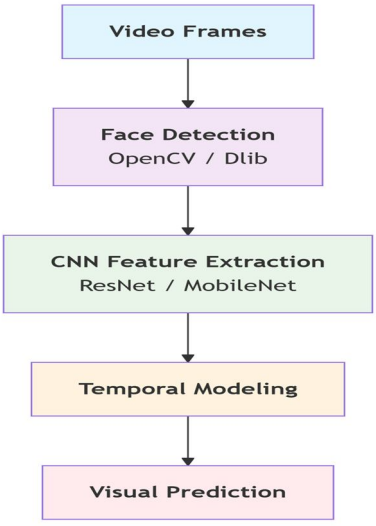


Fig6: visual processing pipeline

A. Advantages of the Proposed System

- Multimodal detection: Simultaneous analysis of text, speech, and facial behavior improves depression detection accuracy.
- High reliability: Fusion of multiple behavioral cues reduces false predictions caused by unimodal limitations.
- Temporal awareness: Sequential and attention-based models capture long-term depressive patterns effectively.
- Non-invasive monitoring: Uses natural behavioral data without requiring intrusive sensors or clinical procedures.
- Scalable design: Modular architecture supports easy integration with telemedicine and mental health platforms.

IV MULTIMODAL FUSION AND CLASSIFICATION

Multimodal fusion integrates complementary information from all modalities to enhance system robustness. The proposed system employs a late fusion strategy, where predictions from individual modalities are combined using a fully connected neural network. Attention mechanisms are incorporated to dynamically weight modality contributions based on contextual relevance.The fused representation is passed through a classification layer to predict depression presence or severity. This approach reduces modality-specific noise and improves generalization across diverse datasets.

V EXPERIMENTAL SETUP AND DATASETS

The system is implemented using Python with TensorFlow, Keras, and OpenCV libraries. Experiments are conducted on benchmark datasets such as DAIC-WOZ and AVEC, which provide synchronized text, audio, and visual data with ground truth depression labels.Data augmentation and normalization techniques are applied to address class imbalance. The model is trained using cross-validation to ensure reliable performance evaluation.

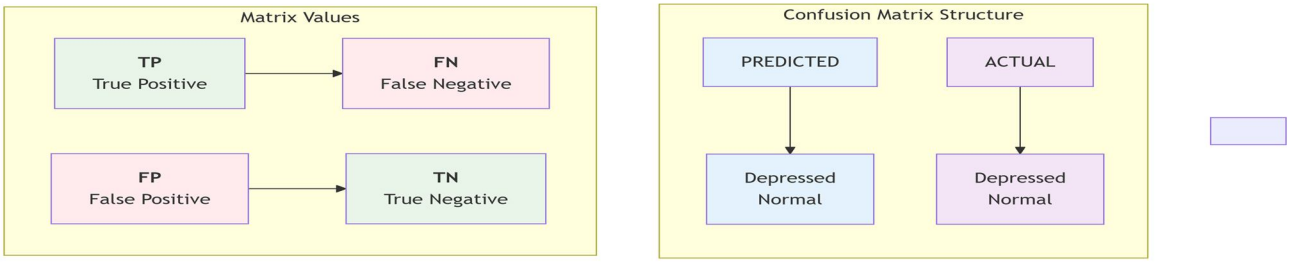


Fig10: confusion matrix

VI CONCLUSIONS

This paper presents a comprehensive AI-based multimodal depression detection framework that integrates textual, acoustic, and visual behavioral analysis. By leveraging advanced deep learning architectures and multimodal fusion strategies, the proposed system addresses key limitations of traditional diagnostic methods and unimodal approaches. The results demonstrate improved accuracy, robustness, and applicability for early depression detection. Future work will focus on explainable AI, privacy-preserving learning, and real-world clinical validation to further enhance system reliability and ethical deployment.

REFERENCES

[1] S. Rasipuram, J. H. Bhat, A. Maitra, B. Shaw, and S. Saha, “Multimodal Depression Detection Using Task-oriented Transformer based Embedding,” in 2022 IEEE Symposium on Computers and Communications (ISCC), Rhodes, Greece, 2022, pp. 1-6. doi:10.1109/ISCC55528.2022.22.9913044.

[2] L. Zhou, Z. Liu, Z. Shangguan, X. Yuan, Y. Li, and B.Hu, “TAMFN: Time-Aware Attention Multimodal Fusion Networkfor Depression Detection,” IEEE Transactions on Neural Systemsand Rehabilitation Engineering, vol. 31, pp. 669–679, 2023. doi:10.1109/TNSRE.2022.3224135.

[3] M. H. Ahmed, Y. Saeed, A. Mehmood, M. Saeed, N. Ahmed, Q. M.Ilyas, S. Iqbal, and N. Abid, “Real-Time Driver Depression Monitoringfor Accident Prevention in Smart Vehicles,” IEEE Access, vol. 12, pp. 79838–79850, 2024. doi:10.1109/ACCESS.2024.3407361.

[4] G. Pranav Arya, G. Ansari, and Y. Saxena, “Multimodal Depression Detection System Using Machine Learning,” in 2023 Second International Conference on Informatics (ICI), Delhi, India, 2023. doi: 10.1109/IC160088.2023.10421362.

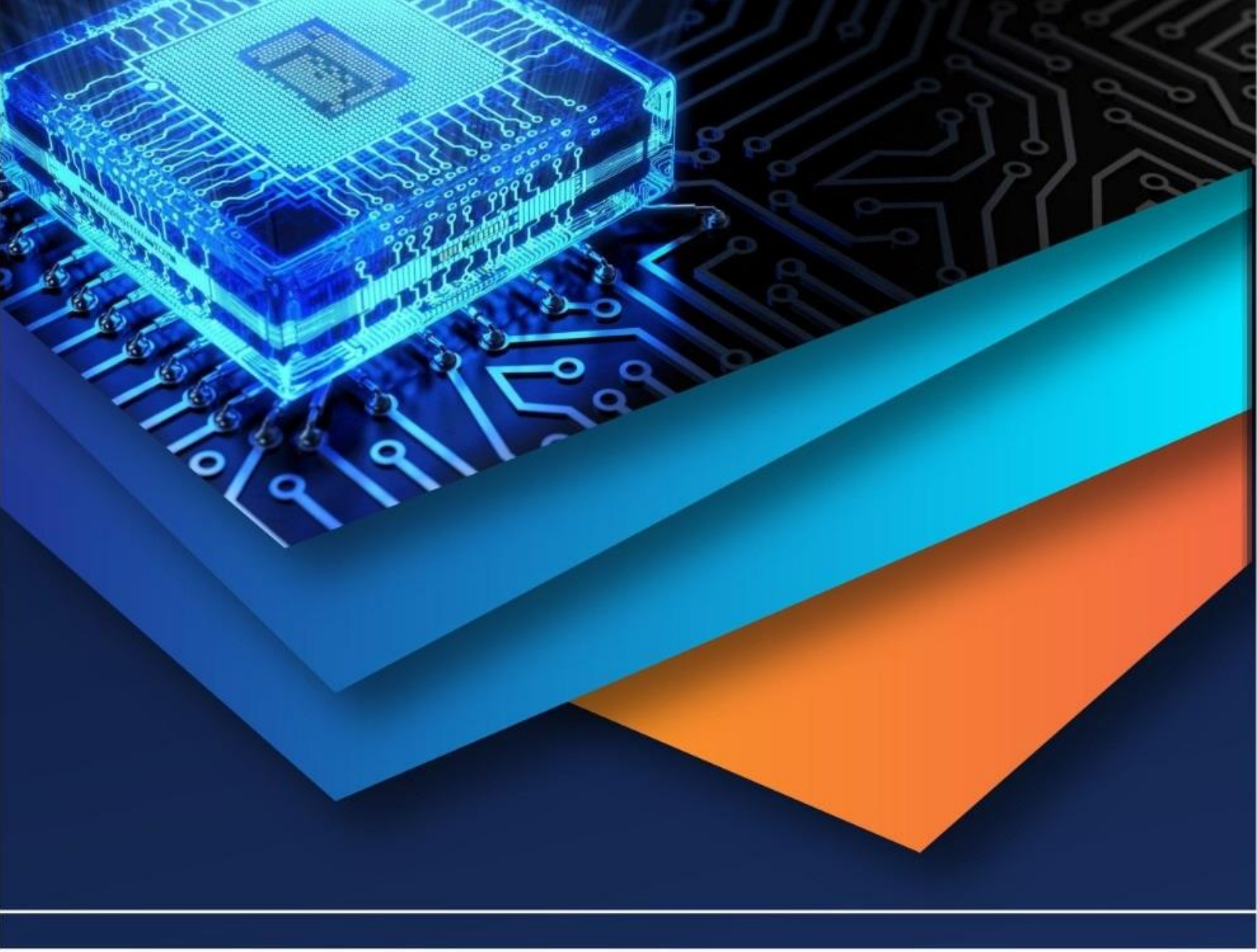
[5] Y. Zhou, X. Yu, Z. Huang, F. Palati, Z. Zhao, Z. He, Y. Feng, and Y. Luo, “Multi-Modal Fused-Attention Network for Depression Level Recognition Based on Enhanced Audiovisual Cues,” IEEE Access, vol. 13, 2025. doi: 10.1109/ACCESS.2025.3545587.

[6] Z. Shangguan, Z. Liu, G. Li, Q. Chen, Z. Ding, and B. Hu, “Dual-Stream Multiple Instance Learning for Depression Detection With Facial Expression Videos,” IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 31, pp. 554, 2023. doi:10.1109/TNSRE.2022.3204757.

[7] M. G. Prasad, “Detection of Depression Using Data Collected from Social Media,” in 2023 International Conference on Network, Multimedia and Information Technology (NMITCON), New Delhi, India, 2023. doi: 10.1109/NMITCON58196.2023.10276373.

[8] C. P. Walia, “Comprehensive Examination of Depression Detection Through Multimedia Content on Social Media Platforms,” in 2025 International Conference on Cognitive Computing in Engineering, Communications, Sciences and Biomedical Health Informatics (IC3ECSBHI), 2025. doi: 10.1109/IC3ECSBHI63591.2025.10990523.

[9] Z. Jiang, Y. Zhou, Y. Zhang, G. Dong, Y. Chen, Q. Zhang, L. Zou, and Y. Cao, “Classification of Depression Using Machine Learning Methods Based on Eye Movement Variance Entropy,” IEEE Access, vol. 12, 2024. doi: 10.1109/ACCESS.2024.3451728.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY