International Journal of Innovative Research in Engineering and Management (IJIREM)
ISSN (Online): 2350-0557, Volume 13, Issue 2, April 2026
DOI: https://doi.org/10.55524/ijirem.2026.13.2.1
Article ID IJIR3172, Pages 1-6
www.ijirem.org

A Multimodal Deep Learning Approach for Depression Detection using Audio and Video Data

T. Srajan Kumar[1], *Kandula Siri Chandana[2], Gavara Archana[3], Gundeti Dhanush Reddy[4] and Jonna Madhu Reddy[5]

[1]Assistant professor, Department of Computer Science and Engineering, Malla Reddy University, Hyderabad, India
[2, 3, 4, 5] B.Tech Scholar, Department of Computer Science and Engineering, Malla Reddy University, Hyderabad, India

Correspondence should be addressed to *Kandula Siri Chandana; kandulasirichandana0@gmail.com

Received: 18 February 2026; Revised: 5 March 2026; Accepted: 17 March 2026

ABSTRACT- Depression is a significant mental health problem in the global society and tends to be undiagnosed because there are no convenient and effective screening procedures. Recognizing depressive symptoms early is important to be able to intervene and treat in time. In this study, the authors suggest a multimodal deep learning-based system of depression detection that takes audio and video inputs to detect patterns of depression. The system evaluates speech features and facial behavioural indicators obtained on the foundation of documented discussions. OpenFace extracts visual (facial action units, gaze direction, and head pose) and speech (processes to extract acoustic aspects of interest in terms of expressing emotions) signals. The model applied in deep learning is the Convolutional Neural Network (CNN) to learn discriminative representations of the extracted features. Audio and visual modalities predicted are then joined together through a fusion mechanism to enhance the overall effectiveness of the classification. The findings of the experiments prove that multimodal analysis is more accurate at detecting depression than unimodal one. The suggested system offers a non-intrusive and scalable system, which can aid mental health practitioners with the initial screening of depression.

KEYWORDS: Depression Detection, Multimodal Learning, Deep Learning, Convolutional Neural Networks, Audio Analysis, facial Behaviour Analysis.

I. INTRODUCTION

The issue of mental health disorders has been an increasing concern in the world over the past few years. In its turn, depression is a condition that impacts millions of people across the globe and has a serious effect on their quality of life [1]. The world of health reports that depression is a major cause of disability and has been attributed to high chances of suicide and other medical ailments.

Conventional diagnosis of depression involves psychological testing and clinical interview done by qualified individuals. These are efficient methods, although they are at times time tricky, subjective and restricted by the availability of mental health specialists. As a result, more interest is developing in the creation of automated systems that can help detect depressive symptoms using behavioral and physiological indicators.

Recent advancements in artificial intelligence and machine learning have enabled the development of automated systems for mental health analysis by interpreting human emotions through speech patterns, facial expressions, and behavioral cues [3] [7]. In this work, a multimodal depression detection system is proposed that analyzes both audio and video signals to identify depressive symptoms. The system extracts facial behavioral features from video recordings and acoustic features from speech signals, which are then analyzed using a deep learning architecture based on Convolutional Neural Networks (CNN) to classify individuals as depressed or non-depressed. By combining predictions from both audio and visual modalities, the proposed approach enhances the accuracy and reliability of depression detection.

II. LITERATURE REVIEW

The recent development in the sphere of artificial intelligence and machine learning promoted researchers to create automated systems of mental health analysis, especially in the direction of depression detection [8]. The analysis of behavioural characteristics is the theme of many studies aimed at the identification of depressive symptoms based on speech, facial expression, and other physiological indicators.

Multiple studies have addressed the issue of speech-based depression detection based on acoustic features derived when voice is recorded [5] [9]. These systems are pitch, tone, speech rate and Mel-Frequency Cepstral Coefficients (MFCC). A research paper has suggested a deep learning framework which makes use of MFCC and spectrogram feature using a Convolutional Neural Network (CNN) to identify depression based on speech samples with high accuracy on standard test sets including DAIC-WOZ and MODMA [4]. The findings established that the speech-based features have the potential of capturing the emotional patterns in depression.

Deep learning networks like CNN, LSTM, and GRU have been studied in other studies regarding depression recognition based on audio signals [13]. To illustrate, a CNN-GRU model was used to analyse mel-spectrograms obtained by speech recordings, which improved the

performance of depression detection by about 20 percent on the DAIC-WOZ dataset [12].

Other than analysis of the speech, facial expression and visual behavioural analysis also has found extensive application in research in depression detection. Emotional states and psychological conditions can be identified with the help of facial expressions, gaze and movements of the head [2] [10].

Other more recent studies have been directed at multimodal systems of depression detection, which can integrate more than one source of data, like audio, video, and text [3] [15], [16]. The multimodal models help garner complementary behavioural cues and tend to be more successful than unimodal methods. Research has also indicated that speech characteristics that are combined with facial expression analysis are more effective in enhancing the detectability [7].

These models have been assessed on several benchmark datasets such as the DAIC-WOZ dataset, which includes recordings of a clinical interview, used to analyse psychological distress [1] [14]. Numerous works use this dataset to train and test depression detection systems based on deep learning approaches.

Despite the large improvements that have been achieved, the current systems still have certain challenges, including insufficient datasets, differences in human emotional expression, and challenges in perceiving the behavioural cues correctly. Thus, the field of multimodal deep learning frameworks combining both audio and visual data is an ongoing research topic to enhance the precision and dependability of automated systems of detecting depression.

III. SYSTEM ARCHITECTURE AND METHODOLOGY

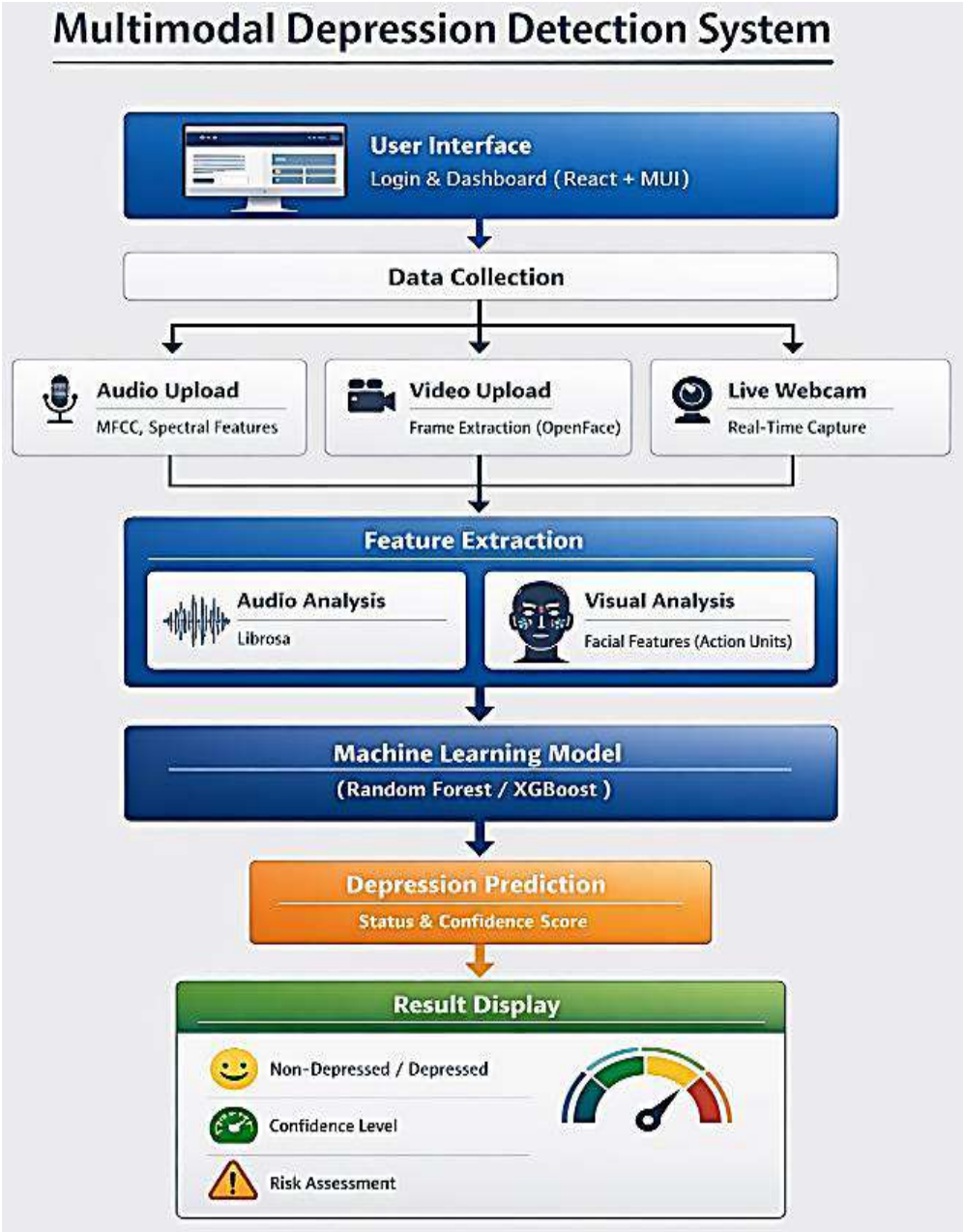

Figure 1: Multimodal Depression Detection System Architecture

The proposed Depression Detection System (see the above figure 1) has a three-layered design: the frontend is a user-friendly web-based app, the middle layer is a machine learning processing layer, which will analyze audio and video and be connected to a backend service that will handle the extraction of the features, model prediction, and the transmission of the data. This layered structure will maintain scale, modularity and effective processing of multimedia data as well as offering a user-friendly experience. The system enables the user to post audio or video files where the user then undergoes an analysis using deep learning techniques to identify whether the person is experiencing depressive symptoms.

A. Architecture Overview

Table 1: Architecture Overview

Component	Technology	Role
Frontend	React.js, HTML,CSS, JavaScript	User interface for uploading audio/video files and displaying prediction results.
Visual Analysis Service	OpenFace, OpenCV	Extract facial features such as facial action units, head pose, and eye gaze from video.
Audio Analysis Service	Librosa, Python	Extract acoustic features such as MFCC, pitch, and speech energy.
Machine Learning Model	CNN,Scikit-learn	Depression classification using extracted multimodal features.
Backend Service	Flask, Python	Handles feature extraction, model inference, and communication with frontend.

In the above table 1, the frontend application is built with React.js and the current web technologies to ensure an interactive interface to the users. With the interface, users are able to post audio records or video files to be analyzed to understand depression. Frontend receives file uploads and shows prediction results, as well as feeds back to the user. The interface can be used in various devices and it is easy to use making it responsive.

Machine Learning Service Layer is the heart of the intelligence of the system. It takes the extracted audio and visual characteristics and inputs them into a deep learning model reliant on the Convolutional Neural Networks (CNN). CNN model acquires the patterns in facial expressions and speech characteristics that can suggest depressive behavior. The model gives a forecast of the input in terms of being a depressed or non-depressed mood and a confidence rating.

The backend service layer is used to control the communication between the frontend interface and the machine learning models. It accepts uploaded files which is then processed by feature extraction modules and the extracted features are sent to the trained model to be classified. Flask and Python are used to implement the backend and ensure the efficient processing of the multimedia data and easy integration of the machine learning libraries.

B. Multimodal Feature Extraction and Depression Detection Module

The detection module of depression is the focal point of the system. It compares facial behavior and speech indicators to determine trends in terms of depressive symptoms. To perform visual analysis, the system works with the video frames, and facial behavioral characteristics are detected with the help of OpenFace. These characteristics are facial action units, eye gaze direction, head pose and facial landmark coordinates. These signs of behavior can be used to capture facial expressions and movements that are related to emotional conditions.

In the audio case, the system uses signal processing methods to obtain acoustic features of speech signals. Significant characteristics like Mel-frequency Cepstral Coefficients (MFCC), variations in pitch, speech power and the rate of speech are calculated. Such characteristics assist in recording the vocal features that can be an indicator of an emotional or psychological state. Audio and visual features are subsequently extracted and a deep learning model that is trained on a Convolutional Neural Network (CNN) are processed. The CNN extracts significant feature representations out of the input data and carries out a classification to show whether the person portrays the characteristic of being depressed. The model generates a prediction label and confidence score of the likelihood of depression.

C. Prediction and User Interaction Module

The prediction module is used to combine the findings of the audio and visual analysis modules and arrive at a final depression detection result. Once the uploaded data has been processed, the prediction is returned to the frontend interface through the backend.

The results of the prediction are then shown on the web interface, which are the classification label (depressed or non-depressed) and the confidence score of the model. This will enable the users to grasp the result of the analysis in a short period. The proposed system will offer an automated technology of detecting depression by combining multimodal data processing and a convenient web interface. The system shows how AI and deep learning can be used to determine possible mental illnesses based on the behaviors of the individual.

IV. IMPLEMENTATION DETAILS

The main idea of the offered system is applied to practice with the help of machine learning and deep learning models combined with a web-based app. The logic of detecting depression is created on the basis of multimodal data analysis that acts on both audio and video data. The system is able to extract the behavioral features of the face of video and the features of speech of the audio. These characteristics are then processed by a deep learning algorithm that is a Convolutional Neural Network (CNN) to determine whether the input labels the depressive symptoms. In this section, one can identify the key functional parts of the system and how they are implemented.

A. Audio / Video Upload and Preprocessing

The input processing module is the audio and video files input by the users who used the web interface. The system

initially checks the file format and prepares a file when a user uploads a file. In the case of video inputs, the system will be able to extract single frames and process them to analyze facial expressions. The frames are rescaled and equalized so that the framework is consistent. OpenFace is then used to extract the facial landmarks and facial behavioral features.

In the case of audio inputs, this system will remove the speech sample in the uploaded file. Some of the steps involved in the preprocessing of the audio signal include noise reduction, normalization and segmentation. These preprocessing procedure steps make sure that the features that have been extracted are befitting model prediction and as well as they fit with the data that have been computed during the training of that model.

B. Visual Feature Extraction and Analysis

The visual analysis module is the one who retrieves facials behavioral features of video records. The system analyses video frames to identify facial features and behavioral changes that can be used to identify emotional states. The system positioning obtains various facial features using OpenFace such as facial action units, eye gaze direction, and movements of the head pose. The features that are sensitive to facial expressions and behavioral trends which more often are linked to emotional and psychological states. The obtained visual features are summed up among the video frames and transformed to feature vectors indicating the behavior of the face of a person in the recording. These vectors of features are then fed into the depression predictor model.

C. Audio Feature Extraction and Speech Analysis

The audio analysis module takes speech signals to find out acoustic features that correspond to emotional and mental states. The system calculates various speech features after extracting the audio in the uploaded file.

Audio processing libraries are used to extract important acoustic features including Mel-Frequency Cepstral Coefficients (MFCC), pitch variation, speech energy and speaking rate. These characteristics assist in recording the vocal patterns which might resemble the depressive behavior like decreased pitch range, slow speech rate or low vocal energy. The extracted audio characteristics are represented in numerical feature vectors and the input of the machine learning model that detects depression.

D. Depression Detection and Prediction.

The most important element of the system is the depression detection module. It takes the vectors of audio and visual features that have been extracted and makes a prediction as to whether the input is a depressed or a non-depressed state. The CNN model is trained based on the predictive characteristics of the input data and determines patterns of depressive behavior. In the prediction process, the feature vectors are fed to many layers of the neural network where the significant patterns are identified. The last step of the model results in an output of classification, which is the predicted mental state and a confidence score to depict the likelihood of depression. Results of the predictions are then forwarded to the frontend application where the user is presented with it.

V. RESULTS AND DISCUSSION

The Multimodal Depression Detection System suggested was tested on audio-visual data as a combination of speech and facial behavioral characteristics. The analysis aimed at the classification performance, comparison of unimodal and multimodal methods, and the usability of the system. System effectiveness was measured using both quantitative measures and qualitative observations.

A. Model Performance Analysis

Training the depression detection model was done on the extracted visual features (facial action units, head pose, and eye gaze) and audio features (MFCC, pitch variation, and speech energy). The sample size was 60 samples (30 of each gender, depressed and non-depressed). Standard classification measures like accuracy, precision, recall, and F1-score were calculated to measure the performance of the model (see the table 2 and figure 2)

Table 2: Visual Model Metrics

Metric	Value
Accuracy	71%
Precision	69%
Recall	72%
F1-Score	70%

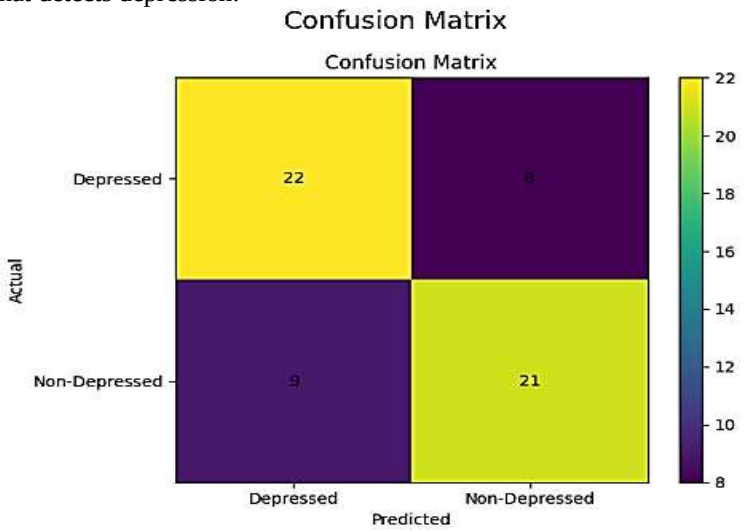

Figure 2: Confusion Matrix of Depression Detection Model

B. Multimodal Depression Evaluation

In order to assess the performance of multimodal learning, the effectiveness of the audio-only, visual-only, and the mixed models were compared.

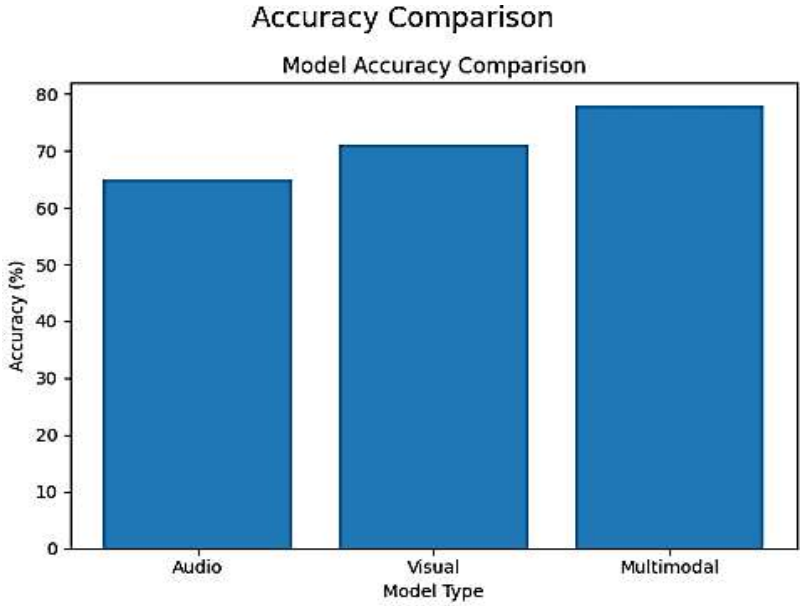

Figure 3: Accuracy Comparison of Audio, Visual and Multimodal Models

In the above figure 3, Multimodal model was the most accurate in which it recorded a score of 78 as compared to audio-only and visual-only models. This shows that the combination of speech and facial features enhances depressive detection because of the role of complementary behavioral information.

C. Analysis of Prediction Output

The system produces a label of classification and a confidence score (see the below table 3).

Table 3: Classification and a confidence score

Sample	Prediction	Confidence Score
S1	Depressed	0.82
S2	Non-Depressed	0.76
S3	Depressed	0.79

The confidence scores are used to describe how likely the model is to predict depression, which therefore makes the prediction of the models better interpretable.

D. Comparative Analysis

Table 4: Comparative Analysis between traditional & proposed system

Feature	Traditional Methods	Proposed System
Depression Assessment	Clinical interviews	Automated AI-based detection
Accuracy	Subjective	Consistent (78%)
Time Required	High	Instant
Behaviour Analysis	Limited	Automated multimodal
Accessibility	Low	High

In the above table 4, we show the comparative analysis between traditional and proposed system. We found that the suggested system offers an alternative system, which is more scaled and faster in comparison to the traditional clinical methods.

E. User Experience and Hands-On Usability

The proposed system was tested in terms of usability with the help of the web-based interface written in React. With the help of the application, a user can upload audio or video recordings and obtain the results of the prediction within several seconds. The interface is easy, interactive and user friendly as it shows the classification result and confidence scores in clear format.

VI. CONCLUSION AND FUTURE WORK

The Multimodal Depression Detection System presented shows that artificial intelligence and deep learning may be applied to aid mental health analysis. The system combines audio and video information to determine behavioral patterns related to depressive symptoms, based on facial expressions and speech features. Facial behavioral patterns are produced with the help of OpenFace, whereas acoustic ones are derived with the help of speech signals. A CNN is used to process these features to classify the people as either depressed or not. It is an application-based system that is deployed as a web-based app in which users are allowed to post audio or video recording and get the results of the predictions accompanied by the confidence score.

The areas of future work will focus on several improvements:

- Mobile Application Development: The system can be created in mobile version so that users can capture speech or facial expression directly through cameras and microphones of smart phones.

- Expanded Dataset and Better Model Training: Enhancing the accuracy and generalization ability of the deep learning model by adding more intense audiovisual records can enhance it.
- Real-Time multimodal Analysis: The system can be more applicable in real-time mental health monitoring use through the introduction of real-time audio and video processing into future versions of the system to analyze behavioral cues in real-time during live interactions, and thus become more appropriate to the related use cases.
- Clinical Decision Support Systems Integration: The system can be combined with the digital healthcare platforms and clinical decision support systems to help mental health professionals analyze the data about patient behavior.
- Explainable Artificial Intelligence: Explainable AI methods can be used in future work to demonstrate which part of the speech or face affects the predictions of the model.
- The proposed multimodal depression detection system is developed as a scalable framework that can be expanded in future on mental health research and health care usages.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] J. Gratch, R. Artstein, G. M. Lucas, *et al.*, "The Distress Analysis Interview Corpus of Human and Computer Interviews," in *Proc. Int. Conf. Language Resources and Evaluation (LREC)*, 2014, pp. 3123–3128. Available from: https://tinyurl.com/343abt3c

[2] Pampouchidou, O. Simantiraki, C. Vazakopoulou, *et al.*, "Automatic Assessment of Depression Based on Visual Cues: A Systematic Review," *IEEE Trans. Affective Computing*, vol. 10, no. 4, pp. 445–470, 2019. Available from: https://doi.org/10.1109/TAFFC.2017.2724035

[3] Z. Yang, X. Xia, and L. Zhao, "Multimodal Depression Detection Using Audio, Visual and Textual Features," *IEEE Access*, vol. 8, pp. 194898–194907, 2020.

[4] S. Al Hanai, M. Ghassemi, and J. Glass, "Detecting Depression with Audio/Text Sequence Modeling of Interviews," in *Proc. Interspeech*, 2018, pp. 1716–1720. Available from: https://www.isca-archive.org/interspeech_2018/alhanai18_interspeech.pdf

[5] F. Eyben, K. R. Scherer, B. W. Schuller, *et al.*, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Trans. Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016. Available from: https://doi.org/10.1109/TAFFC.2015.2457417

[6] T. Baltrusaitis, P. Robinson, and L. P. Morency, "OpenFace: An Open-Source Facial Behavior Analysis Toolkit," in *Proc. IEEE Winter Conf. Applications of Computer Vision (WACV)*, 2016, pp. 1–10. Available from: https://doi.org/10.1109/WACV.2016.7477553

[7] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks," *IEEE J. Selected Topics Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017. Available from: https://doi.org/10.1109/JSTSP.2017.2764438

[8] S. Dham, A. Gupta, and R. Bansal, "Depression Detection Using Machine Learning Techniques on Audio and Visual Data," *Procedia Computer Science*, vol. 167, pp. 2313–2321, 2020.

[9] Y. Yang, C. Fairbairn, and J. F. Cohn, "Detecting Depression Severity from Vocal Prosody," *IEEE Trans. Affective Computing*, vol. 4, no. 2, pp. 142–150, 2013. Available from: https://doi.org/10.1109/T-AFFC.2012.38

[10] H. Dibeklioglu, Z. Hammal, and J. F. Cohn, "Dynamic Multimodal Measurement of Depression Severity Using Facial Expressions and Head Movements," *IEEE J. Biomedical and Health Informatics*, vol. 22, no. 2, pp. 526–536, 2018. Available from: https://doi.org/10.1109/JBHI.2017.2676878

[11] S. Cummins, M. Scherer, J. Krajewski, *et al.*, "A Review of Depression and Suicide Risk Assessment Using Speech Analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015. Available from: https://doi.org/10.1016/j.specom.2015.03.004

[12] B. W. Schuller, S. Steidl, A. Batliner, *et al.*, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proc. Interspeech*, 2013, pp. 148–152. Available from: https://mediatum.ub.tum.de/doc/1189705/document.pdf

[13] Gideon, S. Khorram, Z. Aldeneh, *et al.*, "Progressive Neural Networks for Transfer Learning in Emotion Recognition," in *Proc. ACM Multimedia Conf.*, 2017, pp. 1097–1105. Available from: https://arxiv.org/abs/1706.03256

[14] M. Valstar, J. F. Cohn, T. Baltrusaitis, *et al.*, "AVEC 2016 – Depression, Mood, and Emotion Recognition Workshop and Challenge," in *Proc. ACM Int. Conf. Multimedia*, 2016, pp. 3–10. Available from: https://dl.acm.org/doi/abs/10.1145/2988257.2988258

[15] S. SJ, E. A. Suvi, R. J. Jajin, A. Jagan, and S. PS, "Sign language recognition using deep learning: A systematic review of models and approaches," *International Journal of Innovative Research in Computer Science and Technology (IJIRCST)*, vol. 14, no. 1, pp. 34–42, 2026 Available from: https://doi.org/10.55524/ijircst.2026.14.1.5

[16] G. Tiwary, S. Chauhan, and K. K. Goyal, "Multimodal depression detection using audio visual cues," in *2023 International Conference on Computer Science and Emerging Technologies (CSET)*, Bangalore, India, 2023, pp. 1–5, Available from: https://doi.org/10.1109/CSET58993.2023.10346770