Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 6, December 2026, pp. 5003~5011
ISSN: 2302-9285, DOI: 10.11591/eei.v15i6.11413

An approach to detect depression through sentiment analysis of speech and facial expressions

Uma Yadav[1], Priya Dasarwar (Maidamwar)[2], Shweta V. Bondre[1]
[1]School of Computer Science and Engineering, Ramdeobaba University, Nagpur, India
[2]Department of Computer Science and Engineering, Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International (Deemed University), Pune, India

Article Info

Article history:

Received Sep 17, 2025
Revised Sep 10, 2026
Accepted Oct 4, 2026

Keywords:

Audio features
Deep learning
Depression detection
Facial features
Video extraction

ABSTRACT

Depression is one of the most common mental illnesses among all people, and the early detection of such disease is very hard due to subjective way of detection in classical methods. Audiovisual clues can be used as a source for the detection of depressive symptomatology due to the fact that they are not subjectively estimated. This paper presents the multimodal system for the automatic detection of depressive symptomatology based on audiovisual data analysis. The suggested method includes the following stages: preprocessing, audiovisual feature extraction, dimensionality reduction using factor analysis, multimodal fusion, and classification using convolutional neural network (CNN) and deep belief network (DBN). Two different types of multimodal fusion techniques, namely feature level blend and verdict level blend. The accuracy of the suggested methods was measured in order to investigate the efficiency of the combination of audiovisual data. The results show that feature level blend achieves accuracies of 86.4% and 85.1% with CNN and DBN, respectively, and verdict level blend achieves higher accuracies of 88.2% and 87.6%.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Uma Yadav
School of Computer Science and Engineering, Ramdeobaba University
Nagpur, India
Email: uma.yadav12@gmail.com

1. INTRODUCTION

Depression is one of the most common psychological disorders globally, and its early diagnosis still proves to be problematic because of the subjective nature of classical methods of depression identification including psychological questionnaires and interviews. Early symptoms may not be easily detectable by classical means resulting in inappropriate delay of necessary steps. Recent researches prove that non-verbal features like facial expressions and speech features can be useful in depression detection [1]. Patients with depression tend to have low expressiveness, monotone voice, and longer pauses [1]. Thus, audiovisual automatic analysis can serve as an efficient tool to enhance depression diagnosis [2]. Currently, there are some studies on depression detection which consider multimodal analysis involving speech and facial information in depression detection using machine and deep learning methods. Datasets like Audio/Visual Emotion Challenge (AVEC)-2013, AVEC-2014, and DAIC-WOZ have become widespread sources to test convolutional neural networks (CNNs), recurrent neural networks (RNNs) and other learning algorithms for depression-related features' extraction and classification [3], [4]. However, many problems including selection of efficient audiovisual features, feature dimension reduction, improvement of classification results and others still stay topical [5]. In order to overcome the above issues, this work attempts to explore the use of a multimodal depression detection system based on the combination of video and audio features. This

involves the investigation of feature level blending and verdict level blending approaches based on CNN and deep belief network (DBN) classifiers. The goal here is to determine an optimal fusion and classification approach for depression detection [6]. Research has been conducted recently regarding deep learning and multimodal techniques for automatic depression detection.

He *et al*. [7] introduced deep learning-based global and local attention (DLGA)-CNN, a video-based depression detection system using CNNs along with the attention model. This system is made up of two models called global attention (GA)-CNN and local attention (LA)-CNN, where GA-CNN extracts the global pattern of the face, whereas LA-CNN extracts the local pattern of the face. The proposed model is tested using the AVEC-2013 and AVEC-2014 databases. On the other hand, Shah *et al*. [8] proposed a video-based depression detection technique that uses visual geometry group 16-layer network (VGG-16) CNN architecture. Pampouchidou *et al.* [9] also explored depression and anxiety detection based on videos by using different visual descriptors such as motion history images (MHI), histogram of oriented gradients (HOG), local binary pattern (LBP), and facial geometry feature extraction by means of transfer learning.

Such approaches highlighted the importance of using information about the appearance, movement, and geometry of faces to detect emotions. In multimodal depression prediction, the study [10] proposed a multilevel attention-based network, which used information from several modalities. The main idea of this method was the identification of the important features in different modalities. Automated detection of depression based on multimodal fusion and facial features has been studied in a number of works. Deshpande *et al.* [11] presented research that uses different fusion models and their configurations to analyze the effect of the individual modality on performance and shows good performance results when using different fusion schemes on the datasets AVEC-2013 and AVEC-2014. Dadiz and Ruiz [12] describe the usage of LBP features for construction of the depression classifier using facial videos obtained from the SEMAINE database. Bodenschatz *et al.* [13] suggest a depression classifier construction algorithm based on the video key-frame extraction technique that extracts informative frames from the face. Another popular area of research is the analysis of facial expressions in order to recognize emotions and depression [14], where person-specific active appearance model is used for extraction of the needed facial features. Generally, existing techniques utilize both low-level features that describe visible or audible patterns, and high-level features, that describe emotions and behaviors. Figure 1 illustrates variations in facial expressions used to identify visual cues associated with depressive symptoms.

Although there have been many breakthroughs in automated depression detection, some areas are still lacking in current studies on this topic, particularly those related to multimodal fusion and classifier evaluation. In this study, these limitations have been overcome through the use of video and audio data features along with factor analysis for dimensionality reduction and the use of feature level blend and verdict level blend methods with CNN and DBN classifiers. The verdict level blend with CNN method scored 88.2% accuracy, which is the highest among others.



Figure 1. Facial expression variations considered for identifying visual cues associated with depressive symptoms [15]

2. METHOD

The proposed approach illustrated in Figure 2 conforms to the general methodology suggested by Girard and Cohn [16] and involves multiple sequential steps. First comes the preprocessing step, which is followed by the extraction of the complementary audiovisual features. The next step includes the application of factor analysis to achieve a reduced feature space based on the most relevant information. Finally, the chosen audiovisual features are fused through the use of the multimodal fusion approach, and the result is provided to the classification step where the presence of depressive symptomatology is determined [17]. Algorithm 1 provides an overview of the steps taken for automated detection of depression using multimodal audiovisual data fusion.

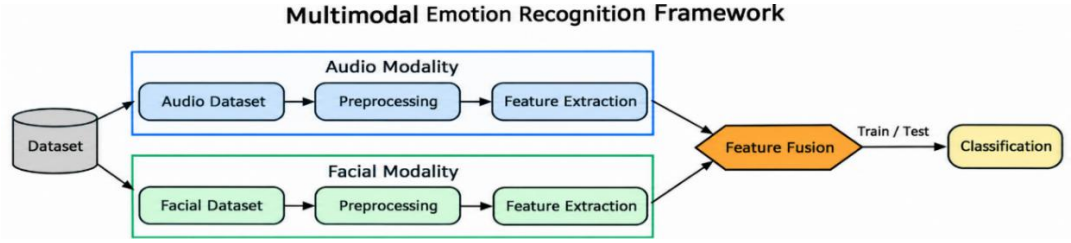


Figure 2. Proposed multimodal depression detection architecture showing the sequential stages of preprocessing, audiovisual feature extraction, dimensionality reduction, fusion, and classification

Algorithm 1. Automated depression detection using multimodal data fusion

Input: Audio and Video Data Input
- Raw video data V
- Raw audio data A

Preprocessing: Preprocessing ensures data quality by addressing noise and enhancing raw inputs.
- Apply noise reduction to V and A to remove unwanted artifacts.
- Enhance video contrast and clarity, and filter audio for improved signal-to-noise ratio.

Feature Extraction: Extract video- and audio-based features to represent behavioral and verbal cues of depression:
- From V: Derive features such as facial expression geometry, motion patterns, and eye movement.
- From A: Extract pitch, tone energy, speech pauses, and prosodic elements.

Dimensionality Reduction: Reduce feature vector complexity to improve computational efficiency and classifier performance: Use Factor Analysis to identify the most significant features in VF (video) and AF (audio).

Fusion of Multimodal Data: Combine video and audio features to leverage complementary information for improved detection accuracy:
- Feature Level Blend: Fuse VF and AF into a unified feature vector.
- Verdict Level Blend: Independently classify VF and AF, then combine the classification outputs.

Classification: Use advanced machine learning models to classify the fused feature set:
- Train a CNN for video-based features.
- Train a DBN for audio-based features.
- Evaluate performance for both fusion strategies using labeled datasets.

Prediction: For new samples, process input through the trained pipeline:
- Pass V and A through preprocessing, feature extraction, and dimensionality reduction steps.
- Apply the fusion strategy and classify the combined features to predict the depression status D.

Output: Deliver the final prediction:
- D=1: Presence of depressive symptoms.
- D=0: Absence of depressive symptoms.

2.1. Datasets

Two publicly available datasets were chosen to ensure that complementary information about audio and facial emotions is included. As for audio, the Dataset Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ) was chosen. This dataset includes 189 clinical interviews which consist of audio and video recordings and answers to questionnaires. The dataset was created to facilitate analysis of different psychological disorders such as depression, anxiety and post-traumatic stress disorder (PTSD). Interviews include not only verbal but also non-verbal behavior of patients. In the case of the analysis of emotions on the face, the Facial Expression Recognition 2013 (FER-2013) database was employed [18]. This database consists of 35,887 facial images that fall under seven different emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral emotion. The images in the database are characterized by different ages, appearances, poses, and facial expressions; therefore, the database is difficult to analyze for emotion recognition. The database is split into three different sub-databases: original training database (OTD) having 28,709 images, public testing database (PTD) having 3,589 images, and final testing database (FTD) having 3,589 images.

2.2. Preprocessing

To preprocess the video, OpenFace was the tool selected to highlight the main feature of the application i.e., 68 face two-dimensional landmarks. Baltrusaitis *et al.* [19] are the creators of OpenFace, a software for extracting facial landmarks, which is exemplified by the pictured and numbered ones in Figure 3. Vonikakis *et al*. [15] also used similar landmarks in their group happiness measure to find distances between particular landmarks, the presence or absence of which is usually linked to joy. Only those frames that had meaningful landmarks detected by the preprocessing method were considered for the subsequent analyses. As a result of audio downsampling to 16 kHz, the silence parts were removed from the time series. Pitch F0 was extracted in the range of "75-455 Hz" and mapped to the domain [0, 1] so that the dependence on the speaker can be removed.

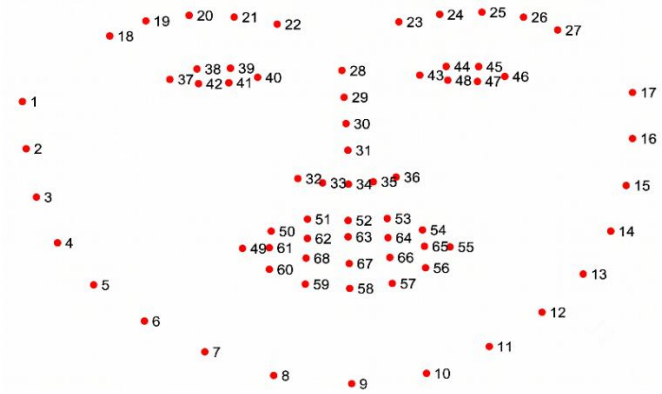
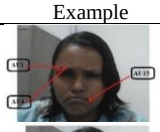


Figure 3. Detection of 68 facial landmarks used for facial feature extraction, including landmarks associated with relevant facial movements such as AU6 and AU12

2.3. Feature extraction

Since smiling is lessened in people feeling sad, this research considers the distances between the landmarks that are affected by smiling [15]. Facial landmarks for the Vera Guth fold and eye blinking are also included. Using the above-mentioned facial landmarks, the selected distances between the landmarks in Figure 3 are considered as time varying features since they change according to the movements of the facial landmarks over time. All distances are normalized using the distance between the lateral facial landmarks (landmark numbers 1 and 17 in Figure 3). Facial expressions are analyzed using the Facial Action Coding System (FACS). This system shows that facial expressions are represented by specific movements of facial muscles. In FACS, the movement of each facial muscle is called an action unit (AU), and specific combinations of AUs can be used to characterize specific emotional expressions. Happiness, for example, is characterized by a combination of AU 6 (cheek raiser) and AU 12 (lip corner puller). Likewise, there are specific combinations of AUs for each type of emotion. Table 1 shows the combination of facial action units (FACS) linked with the chosen emotional expressions. These FACS serve to describe facial expressions and detect visual features related to depression.

Table 1. Different emotions based on AU

Emotion	Description	Example
Sadness	AU1 (inner brow raiser) +AU4 (brow lowerer) +AU15 (lip corner depressor)	
Fear	AU1 (inner brow raiser)+AU2 (outer brow raiser)+AU4 (brow lowerer)+AU5 (upper lid raiser)+AU7 (lid tightener)+AU20 (lip stretcher)+AU26 (jaw drop)	
Anger	AU4 (brow lowerer)+AU5 (upper lid raiser)+AU7 (lid tightener)+AU23 (lip tightener)	
Disgust	AU9 (nose wrinkler)+AU15 (lip corner depressor)+AU16 (lower lip depressor)	

Many statistical features were obtained from the time series data that were extracted, each of which represented the properties of the particular video/audio. The statistical features extracted were the median, mode, mean, range, skewness, standard deviation, variance, kurtosis, entropy, energy, interquartile range, and correlation. For extracting audio features, the Librosa Python library was employed to extract acoustic properties from the recordings. The audio features extracted from the recordings are shown in Table 2.

Table 2. Different audio features

Audio feature	Description
Chroma features	Describes the audio energy distribution over the 12 pitch classes and are often used in the analysis of pitch-related features.
Spectral-roll off	Calculates the frequency below which a certain percentage of the overall spectral energy is distributed and helps in understanding the shape of the power spectrum.
Mel frequency ceptral coefficient	Describes the envelope of an audio signal and are extensively employed in capturing speech features.
Spectral entropy	Indicates the distribution or randomness of the spectral energy and may help in distinguishing between speech, noise, and quiet segments.

2.4. Fusion and classification

The two multimodal fusion techniques considered in this work were feature level blend and verdict level blend. For feature level blend, the extracted video and audio features were first concatenated into one feature vector and passed to the classifier. On the other hand, verdict level blend technique involves processing the video and audio signals separately using different classifiers, and the results are then fused using OR and AND rules. This allows for the use of complementary information from the two modalities in the decision process. In order to classify the data, CNN [20] and DBN [21] models were used. CNN is a deep learning model which is extensively used for visual feature and image classification tasks. In this network, automatic learning of hierarchical features takes place through convolution and pooling, and this makes it possible to extract useful spatial features from the facial features. In our framework, CNN is used for automatic learning of discriminative features from the extracted audiovisual representations in order to detect depression.

3. RESULTS AND DISCUSSION

The suggested strategy to identify signs of depression uses a variety of inputs from video and audio to be combined into a single pipeline of operations that includes: preprocessing, feature extraction, dimensionality reduction, multimodal fusion, and classification. In preprocessing, noise removal and signal enhancement methods are used to increase the quality of raw data. Once preprocessing is done, features specific to each type of modality are extracted. Features related to vision are those that can identify facial expressions and small muscle movements aligned with a person's emotional state, whereas features related to sound are those that describe characteristics of speech, e.g., tone, pitch, and energy levels, which are normally linked to signs of depression. We have used factor analysis for dimensionality reduction to reduce both redundancy and complexity related to computations. This procedure allows us to focus on the most significant latent variables and to drop irrelevant or overly correlated features. The classifiers, which are two deep learning models named CNN and DBN, are then applied to the fused data. Table 3 lists the outcomes of the classifications using these setups.

Table 3. Feature level blend results

Fusion strategy	Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-score (%)
Feature level blend	CNN	86.4	88	84.8	85.27	86.61
Feature level blend	DBN	85.1	87	83.2	83.82	85.38
Verdict level blend	CNN	88.2	90.2	86.2	86.73	88.43
Verdict level blend	DBN	87.6	89.6	85.6	86.15	87.84

The findings reported in Table 3 indicate that the verdict level blend always demonstrated better performance compared to feature level blend for both classifiers. The accuracy of CNN improved from 86.4% to 88.2%, and DBN from 85.1% to 87.6%. Moreover, the accuracy of CNN turned out to be higher compared to DBN in both approaches; the best value of 88.2% was reached with the help of verdict level blend. These findings can be considered evidence that the suggested decision-level fusion setup provided better results for the investigated dataset. Nevertheless, the results may be affected by the size of the dataset, distribution of classes, and individual characteristics of faces and speech. Moreover, misclassification may happen in cases when the sample contains poorly pronounced depressive features or low-quality video

recordings. Thus, despite the good results, further investigations are needed for larger and more diverse sets. To enhance the credibility of the suggested approach, the obtained results were contrasted to the results of recent multimodal depression detection works (2019-2024) and the findings are summarized in Table 4. The analysis shows that the proposed system delivers a good level of performance while resorting to only audiovisual modalities.

Table 4. Comparison with recent multimodal depression detection studies

Study	Year	Modalities	Method	Accuracy (%)
Al Hanai *et al*. [22]	2019	Audio+text	Deep neural network	77
Zhang *et al*. [23]	2023	Audio+video+text	Attention-based fusion	87
Raut *et al*. [24]	2024	Audio+video	MDSTN (3D CNN+late fusion)	74
Hermawan *et al*. [25]	2025	Audio+video	Optimized multimodal DL	87
Proposed method	**2025**	Audio+video	CNN+decision-level fusion	**88.2**

The performance of the developed multimodal depression detection model was also tested by means of confusion matrix and receiver operating characteristic (ROC)-area under the curve (AUC) analysis. Figure 4 demonstrates ROC-AUC curves for both approaches and classifiers. Confusion matrices for feature level blend and verdict level blend approaches using CNN and DBN classifiers are shown in Figure 5.

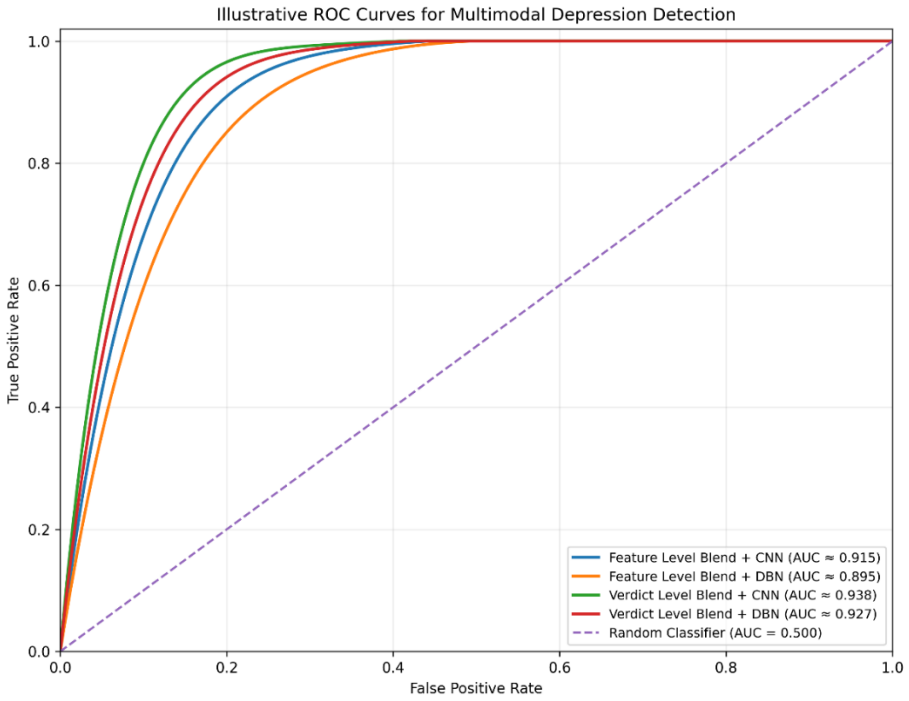


Figure 4. Illustrative ROC-AUC curves comparing the evaluated fusion strategies and classifiers

It can be said that the use of the DAIC-WOZ and FER-2013 datasets gives an adequate experimental basis for testing the proposed multimodal approach, but the lack of diversity in terms of demographic and cultural aspects will make it difficult to generalize the results. Language, cultural, facial, speaking style, and behavioral differences may impact the efficiency of depression detection algorithms. More specifically, FER-2013 describes the general state of facial emotions instead of clinically verified depressive behavior. Thus, the facial information obtained from the dataset can be regarded as additional behavioral evidence, but not depression itself.

However, while CNN delivered the highest accuracy among the proposed models, deployment of the model might face some computational difficulties due to the simultaneous processing of both video frames and audio signals. This might pose some challenges in terms of computational power and memory usage, especially in the context of real-time clinical applications. It is therefore necessary to optimize the model in terms of computation through techniques such as reducing the number of features, efficient neural network architecture, and efficient inference. The goal of this paper was to propose and test a multimodal approach that utilizes the fusion of video and audio data for detecting depressive symptoms. Results highlight

the crucial role of proper fusion techniques and classifier settings in determining recognition performance. The significance of this work lies in showcasing how the integration of various data types can be useful for mental health assessment. This paper mainly focuses on the assessment of model performance through classification accuracy and related assessment methods. Nevertheless, the statistical validation of results through repeated cross-validation and confidence intervals was not done. This is why the actual performance might not fully represent the variance of the model. In future researches, repeated cross-validation, confidence intervals, statistical significance, and cross-dataset tests will be used.

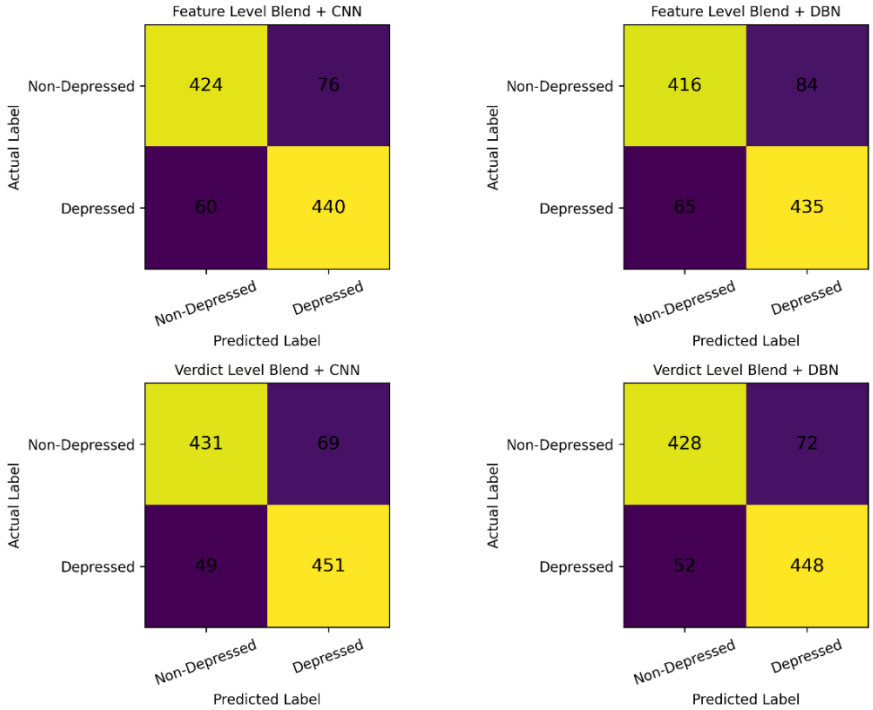


Figure 5. Illustrative confusion matrices for the evaluated fusion strategies and classifiers based on the reported accuracy results

4. CONCLUSION

The present study used a multimodal approach to the detection of depression by utilizing audio and visual modalities via feature level blend and verdict level blend approaches. As can be seen from the results of the experiment, the verdict level blend approach in combination with CNN was able to deliver the best classification accuracy score of 88.2%, demonstrating the viability of decision-level fusion when integrating complementary information coming from audio and visual modalities. The novelty of the methodology of the current study consists in the systematic evaluation of the methods of multimodal fusion when developing automatic depression detection systems. The proposed framework could be potentially applied in telehealth and clinical decision support systems by integrating audio-visual information collected remotely to detect depression signs. However, the proposed system is meant to supplement professional clinical assessment rather than substitute it. There are several limitations in the current study that deserve mention. The datasets used in the study may fail to cover the variety of demographic variables, facial emotions, speech features, and recording scenarios found in real-life situations. Specifically, FER-2013 is a facial emotion recognition database and includes labels related to emotions and not depression-related diagnoses as per clinical criteria. As such, the facial expression data in FER-2013 can only serve as behavioral supportive evidence but not as a measure of depression. These limitations with respect to the datasets might limit the generalizability of the current approach. Thus, the future work in this area will center around the external validation using more diverse and clinically representative datasets gathered in a real-life setting. Repeat experiments, cross-validation, cross-dataset evaluation, and proper statistics will be done to demonstrate the validity and reliability of the results. Further investigation will be done to include additional modalities (physiological and behavioral features) in order to build a broader assessment system. Privacy and fairness aspects will be considered as well. Finally, real-world pilot experiments in telehealth as well as clinical settings will be conducted in order to determine the practical usefulness of the suggested framework. A complete validation,

whether external or clinical, is necessary before the suggested system is deemed operational and ready for deployment in either telehealth or clinical decision-making.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Uma Yadav	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				
Priya Dasarwar (Maidamwar)		✓				✓		✓	✓	✓	✓	✓	✓	
Shweta V. Bondre	✓		✓	✓						✓	✓		✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, [initials: UY], upon reasonable request.

REFERENCES

[1] M. R. Islam, S. Liu, X. Wang, and G. Xu, "Deep learning for misinformation detection on online social networks: a survey and new perspectives," *Social Network Analysis and Mining*, vol. 10, no. 1, p. 82, Dec. 2020, doi: 10.1007/s13278-020-00696-x.

[2] A. Vázquez-Romero and A. Gallardo-Antolín, "Automatic detection of depression in speech using ensemble convolutional neural networks," *Entropy*, vol. 22, no. 6, p. 688, Jun. 2020, doi: 10.3390/e22060688.

[3] J. Park and N. Moon, "Design and Implementation of Attention Depression Detection Model Based on Multimodal Analysis," *Sustainability (Switzerland)*, vol. 14, no. 6, p. 3569, Mar. 2022, doi: 10.3390/su14063569.

[4] P. Wang, B. Yang, S. Wang, X. Zhu, R. Ni, and C. Yang, "Multimodal Depression Detection Network Based on Emotional and Behavioral Features in Conversations," in *Communications in Computer and Information Science*, vol. 1998, 2024, pp. 463–474, doi: 10.1007/978-981-99-9109-9_44.

[5] N. K. Iyortsuun, S. H. Kim, H. J. Yang, S. W. Kim, and M. Jhon, "Additive Cross-Modal Attention Network (ACMA) for Depression Detection Based on Audio and Textual Features," *IEEE Access*, vol. 12, pp. 20479–20489, 2024, doi: 10.1109/ACCESS.2024.3362233.

[6] A. Mallol-Ragolta, Z. Zhao, L. Stappen, N. Cummins, and B. Schuller, "A hierarchical attention network-based approach for depression detection from transcribed clinical interviews," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Sep. 2019, pp. 221–225, doi: 10.21437/Interspeech.2019-2036.

[7] L. He, J. C. W. Chan, and Z. Wang, "Automatic depression recognition using CNN with attention mechanism from videos," *Neurocomputing*, vol. 422, pp. 165–175, Jan. 2021, doi: 10.1016/j.neucom.2020.10.015.

[8] A. Shah, S. Mota, and A. Panchal, "Depression Detection Using Visual Cues," *International Journal of Future Generation Communication and Networking*, vol. 13, no. 3, pp. 1877–1884, 2020.

[9] A. Pampouchidou *et al.*, "Automated facial video-based recognition of depression and anxiety symptom severity: cross-corpus validation," *Machine Vision and Applications*, vol. 31, no. 4, p. 30, May 2020, doi: 10.1007/s00138-020-01080-7.

[10] M. Gavrilescu and N. Vizireanu, "Predicting depression, anxiety, and stress levels from videos using the facial action coding system," *Sensors (Switzerland)*, vol. 19, no. 17, p. 3693, Aug. 2019, doi: 10.3390/s19173693.

[11] Y. Deshpande, S. Patel, M. Lendhe, M. Chavan, and R. Koshy, "Emotion and Depression Detection from Speech," in *Lecture Notes in Networks and Systems*, vol. 154, 2021, pp. 257–265, doi: 10.1007/978-981-15-8354-4_27.

[12] B. G. Dadiz and C. R. Ruiz, "Detecting depression in videos using uniformed local binary pattern on facial features," in *Lecture Notes in Electrical Engineering*, vol. 481, 2019, pp. 413–422, doi: 10.1007/978-981-13-2622-6_40.

[13] C. M. Bodenschatz, F. Czepluch, A. Kersting, and T. Suslow, "Efficient visual search for facial emotions in patients with major depression," *BMC Psychiatry*, vol. 21, no. 1, p. 92, Dec. 2021, doi: 10.1186/s12888-021-03093-6.

[14] Y. An, Z. Qu, N. Xu, and Z. Nima, "Automatic depression estimation using facial appearance," *Journal of Image and Graphics*, vol. 25, no. 11, pp. 2415–2427, 2020, doi: 10.11834/jig.200322.
[15] V. Vonikakis, Y. Yazici, V. D. Nguyen, and S. Winkler, "Group happiness assessment using geometric features and dataset balancing," in *ICMI 2016 - Proceedings of the 18th ACM International Conference on Multimodal Interaction*, Oct. 2016, pp. 479–486, doi: 10.1145/2993148.2997633.
[16] J. M. Girard and J. F. Cohn, "Automated audiovisual depression analysis," *Current Opinion in Psychology*, vol. 4, pp. 75–79, Aug. 2015, doi: 10.1016/j.copsyc.2014.12.010.
[17] U. Yadav and A. K. Sharma, "A novel automated depression detection technique using text transcript," *International Journal of Imaging Systems and Technology*, vol. 33, no. 1, pp. 108–122, Jan. 2023, doi: 10.1002/ima.22793.
[18] R. T. Ionescu, M. Popescu, and C. Grozea, "Local Learning to Improve Bag of Visual Words Model for Facial Expression Recognition," *Workshop on Challenges in Representation Learning, ICML*, pp. 1–6, 2013.
[19] T. Baltrusaitis, C. Ahuja, and L. P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423-443, 2019, doi: 10.1109/TPAMI.2018.2798607.
[20] P. Ray and A. Chakrabarti, "A Mixed approach of Deep Learning method and Rule-Based method to improve Aspect Level Sentiment Analysis," *Applied Computing and Informatics*, vol. 18, no. 1–2, pp. 163–178, Mar. 2022, doi: 10.1016/j.aci.2019.02.002.
[21] Y. Shen, H. Yang, and L. Lin, "Automatic Depression Detection: an Emotional Audio-Textual Corpus and a Gru/Bilstm-Based Model," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing-Proceedings*, 2022, pp. 6247–6251, doi: 10.1109/ICASSP43922.2022.9746569.
[22] T. Al Hanai, M. Ghassemi, and J. Glass, "Detecting depression with audio/text sequence modeling of interviews," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Sep. 2018, pp. 1716–1720, doi: 10.21437/Interspeech.2018-2522.
[23] Z. Zhang *et al.*, "Multimodal Sensing for Depression Risk Detection: Integrating Audio, Video, and Text Data," *Sensors*, vol. 24, no. 12, p. 3714, Jun. 2024, doi: 10.3390/s24123714.
[24] K. Raut, S. Kulkarni, and A. Sawant, "Multimodal spatio-temporal framework for real-world affect recognition," *International Journal of Intelligent Networks*, vol. 5, pp. 340–350, 2024, doi: 10.1016/j.ijin.2024.10.001.
[25] A. Hermawan, B. Daniawan, and J. Nathaniel, "Optimization of Multimodal Deep Learning for Depression Detection," *IJCCS Indonesian Journal of Computing and Cybernetics Systems*, vol. 19, no. 4, pp. 383–394, 2025, doi: 10.22146/ijccs.111407.

BIOGRAPHIES OF AUTHORS



Dr. Uma Yadav is an assistant professor at Ramdeobaba University, formerly Shri Ramdeobaba College of Engineering and Management. She received her B.Tech. and M.Tech. degrees with distinction from Nagpur University in 2009 and 2013, respectively, and holds a Ph.D. in natural language processing and deep learning from GHRU, Amravati. With over 16 years of academic and research experience, her research interests include data science, machine learning, big data management, and data mining. She has authored or co-authored more than 30 research papers in reputed journals and conferences. She can be contacted at email: uma.yadav12@gmail.com.



Dr. Priya Dasarwar (Maidamwar) holds a doctoral degree in Network Security from G H Raisoni University, Amravati. She is currently an assistant professor at Symbiosis Institute of Technology, Nagpur. She received her B.Tech. and M.Tech. degrees with distinction from Nagpur University in 2010 and 2013, respectively. She has published more than 25 research papers in reputed journals and national and international conferences. Her research interests include machine learning, internet of things (IoT), and network security. She can be contacted at email: priyadasarwar21@gmail.com.



Dr. Shweta V. Bondre received her B.Tech. and M.Tech. degrees from Rajiv Gandhi Prudyogiki Vishwavidyalaya, Bhopal, in 2006 and 2013, respectively. Ph.D. thesis submitted at GHRU Amravati. She is an assistant professor at Shri Ramdeobaba College of Engineering and Management. Her research interests include data science, machine learning, cloud computing, big data management, and data mining. She has authored/co-authored 15+ research papers in refereed journals and conferences. She can be contacted at email: bondresv@rknec.edu.