

Depression recognition over fusion of visual and vocal expression using artificial intelligence

Chandan Gautam, Aaradhya Raj, Bhargavee Nemade, Vinaya Gohokar

School of Electronics and Communication Engineering, Dr Vishwanath Karad MIT World Peace University, Pune, India

Article Info

Article history:

Received Dec 29, 2023

Revised Jan 22, 2024

Accepted Feb 8, 2024

Keywords:

CNN

DAIC dataset

Depression scale recognition

LSTM

Visual expression

Vocal expression

ABSTRACT

Depression is a mental illness that usually goes untreated in people and can have catastrophic consequences, including suicidal thoughts. Counselling services are widely available, but because depression is a stigmatized illness, many people who are depressed decide not to seek help. Therefore, it is essential to develop an automated system that can recognize depression in individuals before it worsens. In this study, a novel approach is proposed for identifying depression using a combination of visual and vocal emotions. Long short-term memory (LSTM) is used to assess verbal expressions and convolutional neural networks (CNN) to analyze facial expressions. The proposed system is trained using features of depression from the distress analysis interview corpus (DAIC) dataset and tested on videos of college students with frontal faces. The proposed approach is effective in detecting depression in individuals, with high accuracy and reliability.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Vinaya Gohokar

School of Electronics and Communication Engineering

Dr Vishwanath Karad MIT World Peace University

Pune, India

Email: vinaya.gohokar@mitwpu.edu.in

1. INTRODUCTION

Around 350 million people all over the world suffer from depression, and the frequency is rising every year, according to research on the condition by the World Health Organization (WHO). Epidemiological research found lifetime prevalence of depression to be around 6.9% [1]. Depression is a psychiatric condition that causes extreme sadness and makes it difficult for sufferers to function normally in social situations. More gravely, depression can cause behaviors like self-harm and suicide [2]. Additionally, it may have an impact on significant life events like marriage timing and school success. The majority of persons who receive treatment for depression, according to [3], do not recover from it. The coronavirus disease 2019 (COVID-19) pandemic resulted in extraordinary actions like social segregation, worldwide isolation, and a state of health emergency. This disrupted daily life and caused a significant increase in anxiety and depression symptoms. Over time, the prevalence of depression has more than doubled worldwide. The wellbeing of a person depends on accurate diagnosis and treatment. Traditional diagnostic techniques for depression rely on self-reporting, which can be arbitrary and unreliable, resulting in incorrect diagnoses and insufficient care. Using numerous data sets to create an accurate and dependable detection system could be a way to enhance depression diagnosis and therapy. Currently, the standard clinical approaches for diagnosing depression are primarily based on the clinical diagnoses made by psychiatrists and the patients themselves. The diagnostic and statistical manual of mental disorders (DSM-5), the hamilton depression scale (HAMD), the beck depression inventory (BDI), and the patient health questionnaire (PHQ-9)

are currently the most widely used clinical diagnostic tools. Although correct diagnosis is challenging, these methods have limitations because the diagnostic procedure does not involve objective measurements [4].

A self-administered survey called the beck depression inventory II (BDI-II) [5] is used to gauge how severe depression is in adults and teenagers. Speech and facial expressions are used as biomarkers to determine an individual's level of depression, which can be determined by looking at their BDI-II score as depicted in Table 1. There are 21 multiple-choice questions in the survey, each of which focuses on a different depressive symptom, such as mood, guilt, hunger, or sleep. Based on the person's answers to these questions, the BDI-II score is determined, with each question receiving a number between 0 and 3 that reflects the severity of the symptom. Higher scores on the BDI-II scale, which goes from 0 to 63, indicate more severe depressed symptoms. Mental health practitioners frequently use the BDI-II to diagnose depression, track therapeutic progress, and assess the efficacy of various treatments.

Numerous data sets, including face pictures, and electroencephalogram (EEG), human voice, behaviour, and text, have frequently been employed in research using artificial intelligence (AI) technology to identify depression. Recent background research on the identification of depression shows many techniques and traits used by researchers to identify depression, allowing for significant comparisons between them. Table 2 gives a thorough summary of the state of art literature video data is used as features in [6], [7], AVEC datasets are used by the reserchers and have achieved around 70% accuracy.

Table 1. The BDI-II scores and corresponding depression degree

BDI-II score	Depression degree
0-13	None
14-19	Mild
20-28	Moderate

Table 2. The comparison of the different traits and methods used to evaluate depression

Authors	Features	Dataset	Algorithm	Result	Demerits
Uddin <i>et al.</i> [6]	Video	AVEC2013	CNN,	MAE - 7.04	It relies solely on video data to estimate depression levels.
		AVEC2014	BILSTM	RMSE - 8.93	
Shah <i>et al.</i> [7]	Video	AVEC2014	VGG16	Accuracy - 68.75%	Only utilized facial features.
Singh <i>et al.</i> [8]	Video audio	DAIC - WOZ	PCA, SVM	F1 - 0.82	It does not provide a comparative analysis of its results with other existing methods.
Pampouchidou <i>et al.</i> [9]	Video audio	Own datasets	LBP, HOG	Accuracy - 70%	A small patient group was engaged in the dataset.
Zhou <i>et al.</i> [10]	Video audio	D-Vlog dataset	LSTM, CNN	F1 - 0.75	The dataset was gathered from a particular age group and demographic.
Niu <i>et al.</i> [2]	Audio text	DAIC - WOZ	BLSTM, TCNN	F1 - 0.9074	Potential technical complexities or resource requirements.
Jo and Kwak [11]	Audio text	DAIC - WOZ	BILSTN	Accuracy - 90%	Text data models perform less well than speech data models.
Rustagi <i>et al.</i> [12]	Facial expression and text data	Fer 2013	CNN, RNN	Accuracy - 85%	Only text and visual traits were considered.
Shekar and Kumar [13]	Face circumference audio	Facial expression dataset from Kaggle	CNN	Accuracy - 64%	Only utilized facial features, not considering head posture or eye movement.
Wang <i>et al.</i> [14]	Audio	DAIC - WOZ	CNN, GAN	MAE - 10.21 RMSE - 7.96	It only focuses on audio features and does not consider other potential sources of information.

Changes in speech patterns, facial expressions, body language, and general behavior are a few of the symptoms of depression. Researchers and computers can detect nonverbal signs and subtle behavioral markers that are frequently missed in conventional text-or survey-based assessments by including audio and video data to the dataset. Voice inflection, speech cadence, hesitations, facial expressions, and body language can all reveal important details about a person's emotional state and aid in the detection of depressive symptoms. The incorporation of audio and video data not only improves the sensitivity and accuracy of algorithms for detecting depression, but also offers a more complete picture of a person's wellbeing, aiding in the understanding of the illness and the development of more accurate diagnoses for clinicians and

researchers. Video and accompanied audio is used by researchers in paper [8]–[10]. Various algorithms including principal component analysis (PCA), support vector machine (SVM), long short-term memory (LSTM), and convolutional neural networks (CNN) are implemented to improve accuracy of recognition [2], [11] have implemented the algorithm based on audio and text analysis. Even though 90% accuracy is claimed the technical complexities are more in these algorithms. Facial expressions and text is used [12], [13] to develop the system. Audio features are only criteria used in [14] to achieve the recognition. The datasets frequently used in the literature include:

- DAIC-WOZ
- AVEC 2013
- AVEC 2014
- SEMAINE
- D-Vlog

A set of psychiatric interviews called the distress analysis interview corpus-wizard of oz (DAIC) is used to evaluate people's mental health. This corpus's DAIC-WOZ subset was created using a robotic agent that can communicate with people and recognize verbal and nonverbal indicators of mental illness. 189 audio and video clips as well as in-depth survey responses are included in the collection. The corpus includes WOZ interviews that were used to record and analyze samples for different verbal and nonverbal traits. The data can be utilized to create automated systems for the early detection and prevention of mental health issues and is a useful resource for researchers studying depression and related disorders. A relatively new area of study involves the analysis of audio and video to determine how people are feeling and how they are feeling by observing video frames, audio, and text [15]–[18].

The AVEC 2013 dataset [19]–[23], commonly known as the “audio-visual depression corpus,” contains 150 movies of people recorded with a camera and microphone while they perform depression-oriented tasks in a human-computer interface. Each recording contains only one person, and multiple recordings have a total of 84 subjects. The material contains recordings of 18 test subjects on three, 31 on two and 34 on one, the duration varies from 50 to 20 minutes and the average is 25 minutes. It takes a total of about 240 hours to watch all the clips. The age of individuals in the data varied between 18 and 63 years, with an average age of 31.5 years and a standard deviation of 12.3 years.

Only two of the original 14 activities-which were chosen due to strong participant agreement-are included in the AVEC 2014 selection. 300 movies, ranging in length from 6 to 4 minutes and 8 seconds, were the product. The assessment of anxiety and sadness is the focus of the two activities, which are each independently recorded. Five pairs of previously unreleased clips were additionally included to replace the few videos that were deemed unfit for the competition. The AVEC 2014 source videos were the same as those used in the AVEC 2013. The two tasks in this category are “Free Form” and “The North Wind,” in which participants recite passages from the German novel “Die Sonne und der Wind” (the sun and the north wind), respectively. SEMAINE corpus detects anxiety using face recognition. The corpus consists of dialogues between individuals and a simulated agent and is designed to explore the natural social cues that appear when interacting with an artificial human or robot. Sensitive artificial hearer (SAH) technology was used for data collection. The corpus contains a total of 95 session files with both audio and video, video only, and separate audio files that can be used for testing, interpretation, and case studies. The D-Vlog dataset [1], [24]–[27] is a collection of films made facial expressions. The length and calibre of the movies vary, and the creators come from various social and mental health backgrounds. These videos have been published on YouTube and are open to the general audience. The information, which consists of more than 500 movies, has been utilised for research to build automatic algorithms to identify depression and anxiety based on speech patterns.

The proposed work focuses on, employment of artificial intelligence to recognize depression scales from fusions of visual and vocal expression. A unique method that combines LSTM for audio analysis with CNN for visual analysis is developed which is a novel approach for identifying depression using a combination of visual and vocal emotions. Useful features and models are detected to increase the precision of depression evaluation and identification. The influence of poor data quality on model performance is also addressed. The remainder of the text is structured as follows: The proposed model is thoroughly covered in section 2. Section 3 of this article discusses the results. The conclusion of our work and future research directions are presented in section 4.

2. PROPOSED METHOD

Through physiological research, it has been demonstrated that people with depression behave differently from people in good health in terms of their speech and facial expression. In order to predict a person's state of depression, our suggested method makes use of a multimodal representation framework that

can integrate both visual and audible expressions. As seen in Figure 1, the proposed method consists of different algorithms for audio- and video-based depression recognition. Input, preprocessing, feature extraction, training, and classification are all processes in the system's conventional pipeline. The CNN algorithm is used to train and classify the audio-based depression identification system while the LSTM method is used to train and classify the video-based depression recognition system.

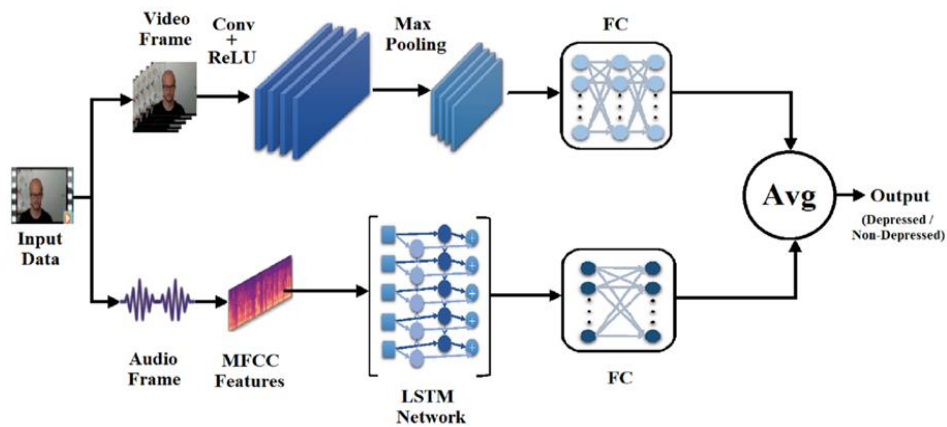


Figure 1. Block diagram for the proposed system

The depression state is decided after combining the output probabilities of the two modalities. The suggested strategy increases the accuracy and robustness of identifying depression by utilizing the advantages of both visual and auditory modalities. Standard performance measures including the confusion matrix, precision, recall, F-measure, and accuracy are used to assess how well the suggested technique performs. Our results demonstrate that the fusion of visual and vocal expressions significantly improves the accuracy of depression detection compared to using only one modality. Proposed CNN technique for diagnosing depression includes the preprocessing, feature extraction, training, and classification phases. The video frames are preprocessed by being scaled to a uniform size and having their pixel values normalized to maintain consistency in the data. The features are extracted using the histogram of oriented gradients (HoG) approach. In the CNN model, which contains 4 convolutional layers, 4 max pooling layers, and 2 fully connected layers, rectified linear units (ReLU) acts as the activation function. The model is trained using backpropagation and a binary cross-entropy loss function. Overfitting is avoided by employing dropout regularization and early halting techniques.

The DAIC-WOZ database is used to diagnose psychological distress disorders such as anxiety, depression, and post-traumatic stress disorder. Before extracting features from the collected visual data for depression detection, pre-processing is necessary. In this step, the video frames are resized and normalized to ensure consistency in the data. Normalization involves adjusting the pixel values so that they have a mean of zero and a standard deviation of one, which helps to reduce any variation caused by differences in lighting or camera settings. These pre-processing steps help to ensure that the data is in a consistent format and ready for feature extraction. A variety of video attributes from the open face framework are retrieved and analyzed. These include,

- Facial landmarks: 68 spots on the face with estimated 2D and 3D coordinates from video
- HOG characteristics for both eyes' gaze direction estimation on the aligned 112×112 region of the face
- Head pose: the 3D posture and orientation of the head Additionally, based on FACET software, assessments of emotion and facial action unit is done. We specifically present the following emotions: "anger," "contempt," "disgust," "disgust," "joy," "fear," "neutral," "sadness," "surprise," "confusion," and "frustration."

Docker is used to expedite and simplify the feature extraction process. By encapsulating the feature extraction algorithms and their dependencies within a Docker container, the consistency and reproducibility of the process across various machines and environments is assured. This approach streamlines the deployment and management of feature extraction workflows, resulting in consistent and reliable outcomes. A training set and a testing set were created from the dataset. The pre-processed video frames are the CNN model's input. The output of the CNN model is a binary classification indicating whether the individual is depressed or not. The CNN model is trained using 4 convolutional layers using ReLU as the activation

function. ReLU, which is applied per pixel and recreates all negative values in the feature map by zero, is an element-wise process. We utilized a flattening layer, two fully linked layers, and a convolutional layer after that. The fully connected layer's purpose is to use these features to categorize the input image into different classes depending on the training dataset. The CNN model then applies a binary cross-entropy loss function and backpropagation. The goal of this loss function is to reduce discrepancy between the predicted and actual classifications, resulting in improved model accuracy. To prevent overfitting, the model is trained with early stopping and dropout regularization techniques. These approaches help improve the model's generalization ability and enhance its performance in processing unseen data.

3. RESULTS AND DISCUSSION

The proposed algorithm was evaluated on a dataset of audio recordings from depressed and non-depressed individuals. It includes preprocessing, feature extraction, data preparation, LSTM network design, training, validation, testing, and post-processing techniques. The preprocessing step involves loading audio signals, removing noise, and normalizing them. The feature extraction step extracts mel-frequency cepstral coefficients (MFCC) features from preprocessed audio signals. MFCC is a widely used feature extraction technique in audio signal processing for various applications such as speech recognition, music classification, and speaker identification. The MFCC feature extraction process involves several steps. The first step is pre-processing, which includes removing any noise or artifacts that may affect the quality of the feature extraction. Common pre-processing techniques include filtering, normalization, and framing. The audio signal is then divided into small, overlapping frames typically 20-30 ms each. Each frame is then multiplied by a window function such as the Hamming or Hann window to reduce spectral leakage. Next, a fourier transform (FT) is performed on each frame to obtain its frequency-domain representation. The frequency-domain representation of each frame is then passed through a mel filter bank, which is a set of overlapping triangular filters that mimic the non-linear frequency response of the human ear. The mel filter bank

3.1. Fusion

The fusion result of the CNN and LSTM models for depression detection is achieved by taking the average of the output from both models. This approach enables the strengths of both models to be combined, resulting in a more accurate and robust prediction. The final prediction is based on a threshold of 0.5, where if the output value is less than 0.5, the result is "not depressed," and if the output value is greater than 0.5, the result is "depressed." This approach provides a binary classification of depression detection that is easily interpretable and actionable for healthcare professionals. By combining the strengths of both models, we can achieve higher accuracy in detecting depression, ultimately improving the diagnosis and treatment of this prevalent mental health disorder.

3.2. Analysis of CNN algorithm

The performance of the proposed depression recognition is evaluated using different optimization algorithms. The CNN and LSTM models are used with each of these optimization algorithms and compared their performance using the test dataset. The results of the CNN algorithm are presented in Table 3.

From the comparative analysis of performance of the Adam and RMSprop optimizer for the training of video-based depression data, it is observed that Adam optimizer outperforms the RMSprop optimizer. The CNN algorithm with Adam optimizer achieved an accuracy of 93%. In the audio-based depression recognition system LSTM algorithm uses the classification of depression and non-depression. The results of audio-based depression recognition using LSTM algorithm with different optimizer are shown in Table 4.

From the comparative analysis of performance of the Adam, RMSprop, stochastic gradient descent (SGD), and Adagrad for the training of audio-based depression data, it is observed that RMSprop optimizer outperform than the other optimizer. The CNN algorithm with RMSprop optimizer achieved an accuracy of 83%. In the testing phase the final decision is given by considering audio as well as video probability of the classes. The average of both the probability is taken and class of maximum probability is considered as output of the depression recognition system. The performance of these models is analyzed using precision, recall, F1-score, and accuracy measures.

Table 3. Analysis of CNN algorithm on video-based depression detection

Algorithm	Optimizer	Precision	Recall	F1-score	Accuracy
CNN	RMSprop	0.96	0.76	0.81	0.83
	Adam	0.93	0.93	0.92	0.93

Table 4. Performance evaluation of LSTM algorithm on training audio data

Algorithm	Optimizer	precision	recall	F1-score	accuracy
LSTM	Adam	0.83	0.76	0.79	0.76
	RMSprop	0.85	0.82	0.83	0.82
	SGD	0.76	0.53	0.58	0.53
	Adagrad	0.03	0.18	0.05	0.18

4. CONCLUSION

The proposed system is an innovative method for identifying depression that incorporates verbal and nonverbal clues. Using CNN for facial expression analysis and LSTM for speech expression processing, accurate classification of individuals as depressed or not depressed is achieved. The suggested approach is evaluated on videos of college students presenting frontal faces after being trained using depression features from the DIAC dataset. The outcomes reveal that proposed method is helpful in identifying depression with high accuracy and dependability. This automated technique is extremely valuable in identifying and treating depression at an early stage. It can lessen the effects of depression on people's lives, such as disturbances to daily routines, academic performance, and social interactions, by offering prompt treatment. There are chances to improve feature extraction at a more sophisticated level in terms of future scope. Investigating additional audio aspects, such as reaction time, pause frequency, and quiet rate, may help us better comprehend the symptoms of depression.





REFERENCES

- [1] K. Mao *et al.*, "Prediction of depression severity based on the prosodic and semantic features with bidirectional LSTM and time distributed CNN," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 2251–2265, Jul. 2023, doi: 10.1109/TAFFC.2022.3154332.
- [2] M. Niu, J. Tao, B. Liu, J. Huang, and Z. Lian, "Multimodal spatiotemporal representation for automatic depression level detection," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 294–307, Jan. 2023, doi: 10.1109/TAFFC.2020.3031345.
- [3] A. Jan, H. Meng, Y. F. B. A. Gaus, and F. Zhang, "Artificial intelligent system for automatic depression level analysis through visual and vocal expressions," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 3, pp. 668–680, Sep. 2018, doi: 10.1109/TCDS.2017.2721552.
- [4] M. A. Wani, M. A. Elaffendi, K. A. Shakil, A. S. Imran, and A. A. A. El-Latif, "Depression screening in humans with AI and deep learning techniques," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 4, pp. 2074–2089, Aug. 2023, doi: 10.1109/TCSS.2022.3200213.
- [5] A. T. Beck, R. A. Steer, and G. Brown, "Beck depression inventory-II," *APA PsycTests*, vol. 78, no. 2, pp. 490–498, 1996, [Online]. Available: <https://psycnet.apa.org/doiLanding?doi=10.1037%2F100742-000>.
- [6] M. A. Uddin, J. B. Joolee, and Y. K. Lee, "Depression level prediction using deep spatiotemporal features and multilayer Bi-LSTM," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 864–870, Apr. 2022, doi: 10.1109/TAFFC.2020.2970418.
- [7] A. Shah, S. Mota, and A. Panchal, "Depression detection using visual cues," vol. IX, no. V, pp. 1–8, 2020.
- [8] N. Singh, R. Kapoor, R. Kapoor, and S. Arora, "Automated major depressive disorder (AMDD) detection using audio-visual feature descriptor and CNN," in *Proceedings - International Conference on Augmented Intelligence and Sustainable Systems, ICAISS 2022*, Nov. 2022, pp. 1456–1459, doi: 10.1109/ICAISS55157.2022.10010907.
- [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] L. Zhou, Z. Liu, Z. Shangguan, X. Yuan, Y. Li, and B. Hu, "TAMFN: time-aware attention multimodal fusion network for depression detection," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 669–679, 2023, doi: 10.1109/TNSRE.2022.3224135.
- [11] A. H. Jo and K. C. Kwak, "Diagnosis of depression based on four-stream model of Bi-LSTM and CNN from audio and text information," *IEEE Access*, vol. 10, pp. 134113–134135, 2022, doi: 10.1109/ACCESS.2022.3231884.
- [12] A. Rustagi, C. Manchanda, N. Sharma, and I. Kaushik, "Depression anatomy using combinational deep neural network," *Advances in Intelligent Systems and Computing*, vol. 1166, pp. 19–33, 2021, doi: 10.1007/978-981-15-5148-2_3.
- [13] P. H. Shekar and K. S. Kumar, "Emotion recognition and depression detection using deep learning," *International Research Journal of Engineering and Technology*, pp. 3031–3036, 2020.
- [14] Z. Wang, L. Chen, L. Wang, and G. Diao, "Recognition of audio depression based on convolutional neural network and generative antagonism network model," *IEEE Access*, vol. 8, pp. 101181–101191, 2020, doi: 10.1109/ACCESS.2020.2998532.
- [15] M. Gheorghe, S. Mihalache, and D. Burileanu, "Using deep neural networks for detecting depression from speech," in *European Signal Processing Conference*, Sep. 2023, pp. 411–415, doi: 10.23919/EUSIPCO58844.2023.10289973.
- [16] H. Solieman and E. A. Pustozarov, "The detection of depression using multimodal models based on text and voice quality features," in *Proceedings of the 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, ElConRus 2021*, Jan. 2021, pp. 1843–1848, doi: 10.1109/ElConRus51938.2021.9396540.
- [17] A. Bailey and M. D. Plumbley, "Gender bias in depression detection using audio features," in *European Signal Processing Conference*, Aug. 2021, vol. 2021-August, pp. 596–600, doi: 10.23919/EUSIPCO54536.2021.9615933.
- [18] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, and B. Schuller, "Hierarchical attention transfer networks for depression assessment from speech," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, May 2020, vol. 2020-May, pp. 7159–7163, doi: 10.1109/ICASSP40776.2020.9053207.
- [19] S. Prabhudesai, M. Parmar, A. Mhaske, and S. Bhagwat, "Depression detection and analysis using deep learning: study and comparative analysis," in *Proceedings - 2021 IEEE 10th International Conference on Communication Systems and Network Technologies, CSNT 2021*, Jun. 2021, pp. 570–574, doi: 10.1109/CSNT51715.2021.9509707.





- [20] Z. Dai, Q. Li, Y. Shang, and X. Wang, "Depression detection based on facial expression, audio, and gait," in *ITNEC 2023 - IEEE 6th Information Technology, Networking, Electronic and Automation Control Conference*, Feb. 2023, pp. 1568–1573, doi: 10.1109/ITNEC56291.2023.10082163.
- [21] C. Fan, Z. Lv, S. Pei, and M. Niu, "Csenet: complex squeeze-and-excitation network for speech depression level prediction," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, May 2022, vol. 2022-May, pp. 546–550, doi: 10.1109/ICASSP43922.2022.9746011.
- [22] H. Kaya, F. Eyben, A. A. Salah, and B. Schuller, "CCA based feature selection with application to continuous depression recognition from acoustic speech features," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, May 2014, pp. 3729–3733, doi: 10.1109/ICASSP.2014.6854298.
- [23] W. C. De Melo, E. Granger, and A. Hadid, "Combining global and local convolutional 3D networks for detecting depression from facial expressions," in *Proceedings - 14th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2019*, May 2019, pp. 1–8, doi: 10.1109/FG.2019.8756568.
- [24] J. Yoon, C. Kang, S. Kim, and J. Han, "D-vlog: multimodal vlog dataset for depression detection," *Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022*, vol. 36, pp. 12226–12234, 2022, doi: 10.1609/aaai.v36i11.21483.
- [25] Y. Tao, M. Yang, Y. Wu, K. Lee, A. Kline, and B. Hu, "Depressive semantic awareness from vlog facial and vocal streams via spatio-temporal transformer," *Digital Communications and Networks*, pp. 2352–8648, Mar. 2023, doi: 10.1016/j.dcan.2023.03.007.
- [26] W. Zheng, L. Yan, and F. Y. Wang, "Two birds with one stone: knowledge-embedded temporal convolutional transformer for depression detection and emotion recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 2595–2613, Oct. 2023, doi: 10.1109/TAFFC.2023.3282704.
- [27] 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.

BIOGRAPHIES OF AUTHORS







Chandan Gautam     holding a B.Tech degree in Electronics and Communication Engineering from MIT WPU in 2023, Pune, is an accomplished Software Developer at Northern Trust. Awarded the first rank in her department and a gold medal, also served as the IEEE Chairperson, MITWPU in 2022. She demonstrates a compelling blend of leadership skills and academic prowess, contributing significantly to advancements in affective computing and multimodal emotion recognition, underscoring her dedication to the field. She can be contacted at email: chandangautam0607@gmail.com.







Aaradhya Raj     with a B.Tech degree in Electronics and Communication Engineering from MIT WPU (2023, Pune), currently works at Harman as a software developer, demonstrating proficiency and dedication to advancing technology. Her research interests encompass speech processing, computer vision, affective computing, and machine learning. She can be contacted at email: aaradhyadiscovery@gmail.com.



Bhargavee Nemade     a B.Tech. in Electronics and Communication graduate from MIT WPU (2023) and Test Engineer at IDEaS-A SAS Company, has made substantial contributions to advancing affective computing and multimodal emotion recognition, highlighting her dedication to innovation at the intersection of technology and emotion analysis. She can be contacted at email: bhargaveenemade@gmail.com.



Dr. Vinaya Gohokar     is working as Professor in School of ECE, MIT World Peace University. She has more than 30 years of experience in teaching and research. Her area of interest includes internet of things (IoT), machine learning, edge intelligence, and computer vision. She has many grants from DST, AICTE to her credit. She has published research papers in various journals and conferences. She can be contacted at email: vinaya.gohokar@mitwpu.edu.in.