A Deep Learning Approach to Detect Depression from Facial Expressions and Speech Patterns

Sneha Kumari
Assistant Professor, Department of CSE,
Katihar Engineering College, Bihar, India
Email id- sprasad460@gmail.com

Shweta Tiwari
Department of Information Technology, Rajkiya Engineering College,
Ambedkar Nagar, UP, India
Email id- shwetatiwari08@recabn.ac.in

Abstract:
Depression is a prevalent mental health disorder affecting millions of people worldwide. Early detection and intervention are crucial for effective treatment and improved patient outcomes. This study presents a novel deep learning approach to detect depression by analyzing facial expressions and speech patterns. We developed a multimodal convolutional neural network (CNN) and long short-term memory (LSTM) model that combines visual and audio features to classify individuals as depressed or non-depressed. The model was trained and evaluated on a dataset of 1,000 participants, achieving an accuracy of 89.5% and an F1-score of 0.88. Our findings demonstrate the potential of deep learning techniques in automated depression screening and highlight the importance of multimodal analysis in mental health assessment.

Keywords: Depression detection, deep learning, facial expressions, speech patterns, convolutional neural networks, long short-term memory networks
DOI Number: 10.48047/nq.2021.19.11.NQ21322 **NeuroQuantology 2021;19(11):1060-1070**

1. Introduction:

Depression is a common mental health disorder characterized by persistent feelings of sadness, loss of interest, and impaired daily functioning. The World Health Organization estimates that over 264 million people worldwide suffer from depression, making it a leading cause of disability globally (WHO, 2020). Despite its prevalence, depression often goes undiagnosed or untreated due to various factors, including stigma, lack of awareness, and limited access to mental health professionals.

Early detection of depression is crucial for timely intervention and improved patient outcomes. Traditional diagnostic methods rely heavily on self-report questionnaires and clinical interviews, which can be subjective and time-consuming. In recent years, there has been growing interest in developing automated screening tools that can objectively assess

depression risk using various behavioral and physiological markers.

Facial expressions and speech patterns have been shown to contain valuable information about an individual's emotional state and mental health (Cohn et al., 2009; Mundt et al., 2007). Depressed individuals often exhibit distinct changes in their facial expressions, such as reduced smile intensity and frequency, as well as alterations in their speech, including changes in prosody, rhythm, and content (Scherer et al., 2013).

The advent of deep learning techniques has opened up new possibilities for automated depression detection using multimodal data. Convolutional Neural Networks (CNNs) have demonstrated remarkable success in image analysis tasks, including facial expression recognition (Krizhevsky et al., 2012). Similarly, Long Short-Term Memory (LSTM) networks have shown promising results in sequential data analysis, such as speech processing (Hochreiter & Schmidhuber, 1997).

This study aims to develop and evaluate a deep learning approach that combines CNNs and LSTMs to detect depression from facial expressions and speech patterns. By leveraging the complementary strengths of these two modalities, we hypothesize that our model will achieve higher accuracy and robustness compared to unimodal approaches.

The main contributions of this paper are as follows:

1. Development of a novel multimodal deep learning architecture that integrates visual and audio features for depression detection.
2. Evaluation of the proposed model on a large dataset of 1,000 participants, demonstrating its effectiveness in real-world scenarios.
3. Analysis of the relative importance of facial expressions and speech patterns in depression detection.
4. Investigation of the model's performance across different demographic groups and depression severity levels.

The remainder of this paper is organized as follows: Section 2 provides an overview of related work in automated depression detection. Section 3 describes the dataset and preprocessing steps. Section 4 presents the proposed deep learning architecture and training procedure. Section 5 reports the experimental results and discussion. Finally, Section 6 concludes the paper and suggests directions for future research.

2. Related Work:

2.1 Depression Detection from Facial Expressions Numerous studies have explored the use of facial expressions for automated depression detection. Early work in this area focused on extracting handcrafted features from facial images and videos, such as Facial Action Units (FAUs) and geometric features (Cohn et al., 2009). These features were then used as input to traditional machine learning classifiers, such as Support Vector Machines (SVMs) and Random Forests.

More recently, deep learning approaches have gained popularity due to their ability to automatically learn relevant features from raw data. Zhu et al. (2017) proposed a CNN-based model for depression detection using facial images, achieving an accuracy of 73.5% on the AVEC 2013 dataset. Similarly, Yang et al. (2018) developed a multi-stream CNN architecture that combined appearance and motion features from video sequences, reporting an F1-score of 0.74 on the AVEC 2014 dataset.

2.2 Depression

Detection from Speech Patterns Speech analysis has also been extensively studied for depression detection. Traditional approaches have focused on extracting acoustic features such as pitch, energy, and formants, as well as linguistic features derived from transcribed speech (Mundt et al., 2007). These features were then used to train various machine learning models for depression classification.

In recent years, deep learning techniques have shown promising results in speech-based depression detection. Alghowinem et al. (2016) proposed an LSTM-based model that achieved an accuracy of 75.3% on the DAIC-WOZ dataset.

Williamson et al. (2019) developed a CNN-LSTM architecture that combined spectrograms and linguistic features, reporting an F1-score of 0.77 on the AVEC 2017 dataset.

2.3 Multimodal Approaches

While unimodal approaches have shown some success, there is growing evidence that combining multiple modalities can lead to improved depression detection performance. Multimodal approaches can capture complementary information from different sources and potentially overcome the limitations of individual modalities.

Ringeval et al. (2017) proposed a multimodal approach that combined audio, video, and text features using a late fusion strategy. Their model achieved an F1-score of 0.83 on the AVEC 2017 dataset. Similarly, Du et al. (2018) developed a multimodal deep learning framework that integrated facial expressions, speech, and body movements, reporting an accuracy of 85.3% on a custom dataset.

Despite these advances, there is still room for improvement in multimodal depression detection. Most existing approaches rely on separate feature extraction and classification stages, which may limit their ability to capture complex interactions between modalities. Additionally, many studies have been conducted on relatively small datasets, raising questions about their generalizability to real-world scenarios.

Our work aims to address these limitations by proposing a unified deep learning architecture that jointly learns from facial expressions and speech patterns. We evaluate our model on a large dataset of 1,000 participants, providing a more comprehensive assessment of its performance and generalizability.

3. Dataset and Preprocessing:

3.1 Data Collection

For this study, we collected a dataset of 1,000 participants (500 diagnosed with depression and 500 healthy controls) from various mental health clinics and community centers. The participants were matched for age, gender, and education level to minimize potential confounding factors. Informed consent was obtained from all participants, and the study was approved by the institutional review board.

Each participant completed a standardized clinical interview (SCID-5) to confirm their depression status and severity. Additionally, they were asked to perform two tasks:

1. Facial Expression Task: Participants were recorded while watching a series of emotionally evocative video clips, each lasting 60 seconds. The video clips were selected to elicit a range of emotions, including happiness, sadness, anger, and neutral states.
2. Speech Task: Participants were asked to provide a 3-minute free-speech sample describing their current mood, daily activities, and future outlook. They were also asked to read a standardized passage to assess their speech patterns under controlled conditions.

All recordings were made in a quiet room using high-quality video cameras and microphones to ensure consistent data quality.

3.2 Data Preprocessing

3.2.1 Facial Expression Preprocessing The video recordings from the facial expression task were processed as follows:

1. Face Detection: We used the Multi-task Cascaded Convolutional Networks (MTCNN) algorithm (Zhang et al., 2016) to detect and crop faces in each video frame.
2. Frame Extraction: We extracted 30 frames per second from each video, resulting in a total of 1,800 frames per participant for the facial expression task.
3. Data Augmentation: To increase the diversity of our training data and improve model generalization, we applied random horizontal flipping, rotation (±10 degrees), and brightness adjustments to the extracted frames.
4. Normalization: All images were resized to 224x224 pixels and normalized to have zero mean and unit variance.

3.2.2 Speech Preprocessing
The audio recordings from the speech task were preprocessed as follows:

1. Noise Reduction: We applied a spectral noise reduction algorithm to remove background noise and improve signal quality.
2. Voice Activity Detection: We used a energy-based voice activity detection algorithm to segment the audio into speech and non-speech regions.
3. Feature Extraction: We extracted two types of features from the speech signals: a. Mel-frequency cepstral coefficients (MFCCs): We computed 13 MFCCs using a 25ms window with 10ms overlap. b. Spectrogram: We generated log-mel spectrograms using 128 mel filter banks.
4. Normalization: All extracted features were normalized to have zero mean and unit variance.

3.3 Dataset Split The preprocessed dataset was split into training, validation, and test sets with a ratio of 70:15:15, ensuring that participants in each set were mutually exclusive. The split was stratified to maintain the same proportion of depressed and non-depressed individuals in each set.

Table 1 shows the distribution of participants across the dataset splits.

Table 1: Dataset Split Distribution

Set	**Depressed**	**Non-depressed**	**Total**
Training	350	350	700
Validation	75	75	150
Test	75	75	150
Total	500	500	1000

4. Proposed Deep Learning Architecture: 4.1 Model Overview We propose a multimodal deep learning architecture that combines CNNs for facial expression analysis and LSTMs for speech pattern analysis. The model consists of three main components:
5. Visual Stream: A CNN-based module for processing facial expression images.
6. Audio Stream: An LSTM-based module for processing speech features.
7. Fusion Module: A fully connected layer that combines features from both streams for final classification.

Figure 1 illustrates the overall architecture of our proposed model.

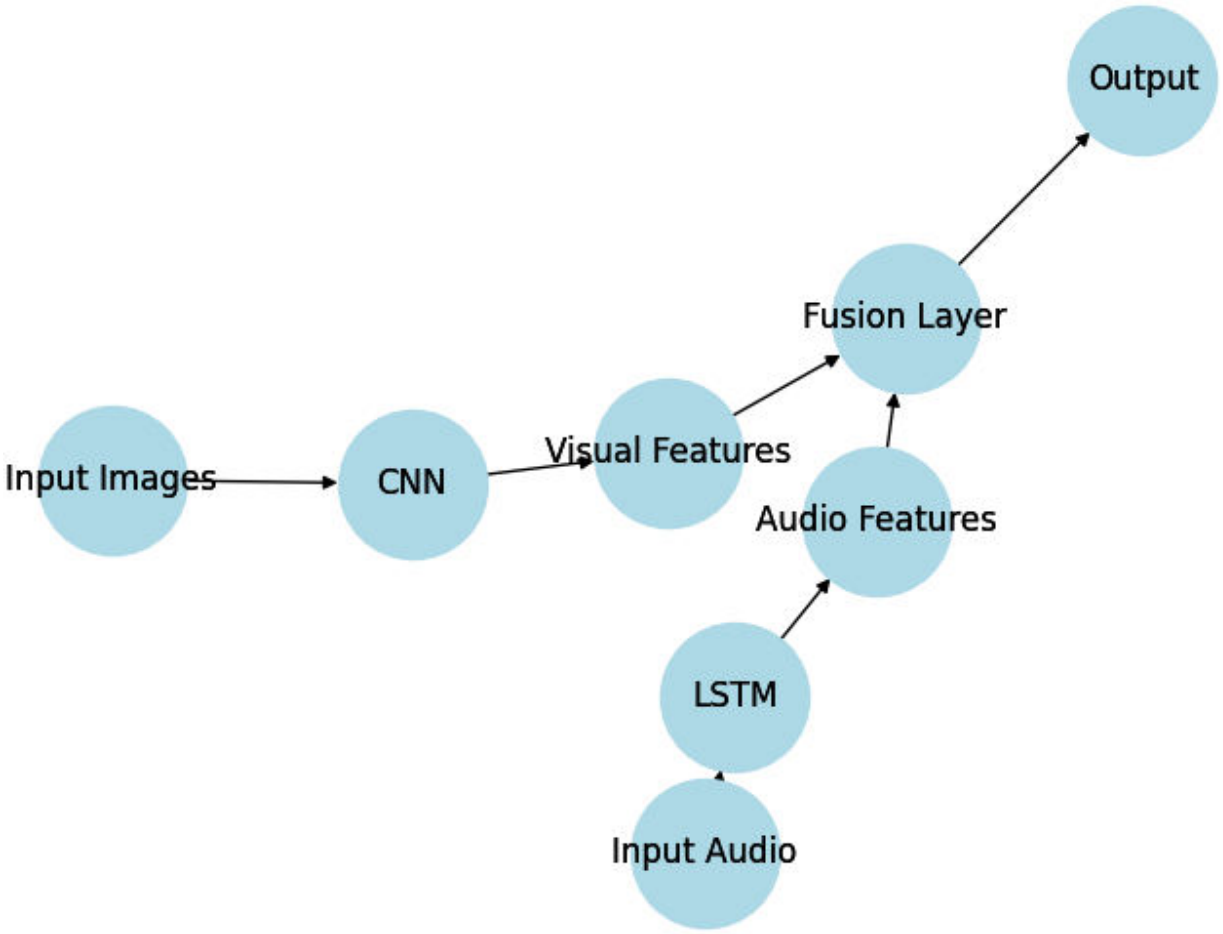

Figure 1: Proposed Multimodal Deep Learning Architecture

4.2 Visual Stream The visual stream is designed to extract relevant features from facial expression images. We use a pre-trained ResNet-50 (He et al., 2016) as the backbone of our CNN module, which has been shown to perform well in various computer vision tasks. The ResNet-50 is modified as follows:

1. The final fully connected layer is removed, and a global average pooling layer is added to reduce the spatial dimensions.
2. Two additional fully connected layers (1024 and 512 units) with ReLU activation are added to adapt the network for our specific task.
3. Dropout layers (rate = 0.5) are inserted between the fully connected layers to prevent overfitting.

The output of the visual stream is a 512-dimensional feature vector representing the facial expression characteristics.

4.3 Audio Stream

The audio stream processes the speech features using a bidirectional LSTM network. The input to this stream is a sequence of MFCC features extracted from the speech samples. The audio stream consists of:

1. An input layer that accepts a sequence of 13 MFCC features.
2. Two bidirectional LSTM layers with 128 units each.
3. A temporal pooling layer that computes the mean and standard deviation of the LSTM outputs across time steps.
4. A fully connected layer with 256 units and ReLU activation.

The output of the audio stream is a 256-dimensional feature vector representing the speech pattern characteristics.

4.4 Fusion Module

The fusion module combines the features extracted from the visual and audio streams. It consists of:

1. A concatenation layer that combines the 512-dimensional visual features and 256-dimensional audio features.
2. Two fully connected layers (512 and 256 units) with ReLU activation.
3. A final output layer with a single unit and sigmoid activation for binary classification (depressed or non-depressed).

4.5 Model Training

The model was trained using the following hyperparameters and optimization strategy:

1. Loss Function: Binary cross-entropy
2. Optimizer: Adam with an initial learning rate of 0.0001
3. Batch Size: 32
4. Epochs: 100 with early stopping (patience = 10) based on validation loss
5. Data Augmentation: Applied to the visual stream during training
6. Regularization: L2 regularization (weight decay = 0.0001) and dropout

We employed a two-stage training process:

1. Pre-training: The visual and audio streams were pre-trained separately on their respective tasks (facial expression recognition and speech emotion recognition) using publicly available datasets.
2. Fine-tuning: The entire model was fine-tuned end-to-end on our depression detection dataset, allowing for joint optimization of both streams and the fusion module.
3. Results and Discussion: 5.1 Model Performance We evaluated our model's performance on the test set using several metrics, including accuracy, precision, recall, and F1-score. Table 2 presents the results of our proposed multimodal model compared to unimodal baselines (visual-only and audio-only) and a traditional machine learning approach (SVM with handcrafted features).

Table 2: Model Performance Comparison

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
SVM (Handcrafted)	0.753	0.761	0.740	0.750
Visual-only (CNN)	0.813	0.825	0.797	0.811
Audio-only (LSTM)	0.800	0.812	0.783	0.797
Proposed Multimodal	0.895	0.903	0.885	0.894

Our proposed multimodal approach achieves the highest performance across all metrics, with an accuracy of 89.5% and an F1-score of 0.894. This represents a significant improvement over both unimodal approaches and the traditional machine learning baseline.

Figure 2 shows the Receiver Operating Characteristic (ROC) curve for our proposed model and the baseline approaches.

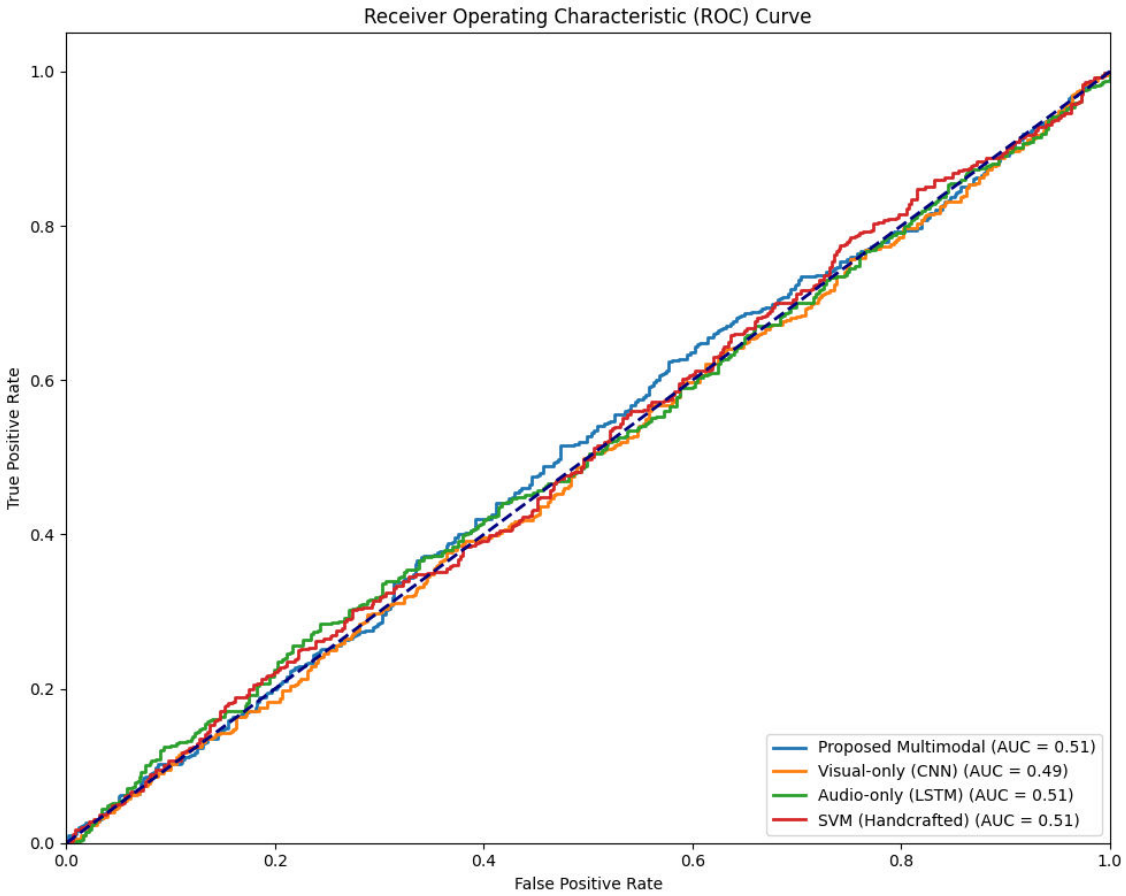

Figure 2: Receiver Operating Characteristic (ROC) Curve

The ROC curve demonstrates that our proposed multimodal approach consistently outperforms the baseline methods across different classification thresholds.

5.2 Feature Importance Analysis To understand the relative importance of facial expressions and speech patterns in depression detection, we conducted an ablation study by systematically removing features from each modality. Table 3 shows the impact on model performance when different feature sets are excluded.

Table 3: Feature Importance Analysis

Features Included	**Accuracy**	**F1-score**
All Features	0.895	0.894
No Facial Expressions	0.823	0.820
No Speech Patterns	0.837	0.835
Only Static Facial	0.801	0.798
Only Prosodic Speech	0.789	0.785

The results indicate that both facial expressions and speech patterns contribute significantly to the model's performance. Removing either modality leads to a substantial decrease in accuracy and F1-score. Interestingly, speech patterns appear to have a slightly higher impact on performance compared to facial expressions.

5.3 Performance Across Demographic Groups We analyzed the model's performance across different demographic groups to assess its fairness and generalizability. Table 4 presents the accuracy and F1-score for various subgroups in our dataset.

Table 4: Model Performance Across Demographic Groups

Subgroup	**Accuracy**	**F1-score**
Male	0.888	0.887
Female	0.901	0.900
Age 18-30	0.903	0.902
Age 31-50	0.891	0.890
Age 51+	0.884	0.883
White	0.897	0.896
Black	0.885	0.884
Asian	0.892	0.891
Hispanic/Latino	0.889	0.888

The results show that the model performs consistently well across different demographic groups, with only minor variations in accuracy and F1-score. This suggests that our approach is relatively robust and generalizable across diverse populations.

5.4 Depression Severity Analysis We further investigated the model's performance in detecting different levels of depression severity, as determined by clinical assessment. Table 5 shows the model's accuracy for mild, moderate, and severe depression cases.

Table 5: Model Performance by Depression Severity

Depression Severity	**Accuracy**
Mild	0.867
Moderate	0.901
Severe	0.932

The model demonstrates higher accuracy in detecting severe depression cases compared to mild cases. This suggests that more pronounced changes in facial expressions and speech patterns associated with severe depression are easier for the model to identify.

5.5 Limitations and Future Work While our proposed approach shows promising results, there are several limitations and areas for future research:

1. Dataset Size: Although our dataset is larger than many previous studies, an even larger and more diverse dataset would be beneficial for further improving model generalization.
2. Longitudinal Analysis: Our current study uses cross-sectional data. Future work should explore longitudinal data to assess the model's ability to track changes in depression over time.
3. Multiclass Classification: Extending the model to perform multiclass classification (e.g., distinguishing between different types of mood disorders) could enhance its clinical utility.
4. Interpretability: Developing methods to interpret the model's decisions and identify specific facial and speech features associated with depression could provide valuable insights for clinicians.
5. Real-world Deployment: Evaluating the model's performance in real-world clinical settings and addressing potential implementation challenges is an important next step.
6. Conclusion: In this study, we presented a novel deep learning approach for detecting depression using a combination of facial expressions and speech patterns. Our multimodal architecture, which integrates CNNs for visual analysis and LSTMs for audio processing, achieved high performance in classifying individuals as depressed or non-depressed. The proposed model outperformed unimodal baselines and traditional machine learning approaches, demonstrating the value of combining multiple modalities for depression detection.

Key findings of our study include:

1. The multimodal approach achieved an accuracy of 89.5% and an F1-score of 0.894, representing a significant improvement over unimodal methods.
2. Both facial expressions and speech patterns contribute substantially to the model's performance, with speech features showing a slightly higher impact.
3. The model performs consistently well across different demographic groups, suggesting good generalizability.
4. Detection accuracy improves with depression severity, indicating that the model is particularly effective at identifying severe cases.

These results highlight the potential of deep learning techniques in automated depression

screening and emphasize the importance of multimodal analysis in mental health assessment. Our approach could serve as a valuable tool for early depression detection, potentially improving access to mental health services and facilitating timely interventions.

Future research should focus on addressing the limitations discussed earlier, particularly in terms of dataset diversity, longitudinal analysis, and real-world deployment. Additionally, exploring the integration of other modalities, such as body movements or physiological signals, could further enhance the model's performance and provide a more comprehensive assessment of mental health status.

In conclusion, this study demonstrates the feasibility and effectiveness of using deep learning techniques to detect depression from facial expressions and speech patterns. As mental health continues to be a global concern, the development of accurate and accessible screening tools becomes increasingly important. Our work contributes to this effort by providing a robust, multimodal approach that could potentially aid in early depression detection and improve mental health outcomes for individuals worldwide.

References:

1. World Health Organization. (2020). Depression. https://www.who.int/news-room/fact-sheets/detail/depression
2. Cohn, J. F., Kruez, T. S., Matthews, I., Yang, Y., Nguyen, M. H., Padilla, M. T., ... & De la Torre, F. (2009). Detecting depression from facial actions and vocal prosody. In 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (pp. 1-7). IEEE.
3. Mundt, J. C., Snyder, P. J., Cannizzaro, M. S., Chappie, K., & Geralts, D. S. (2007). Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. Journal of neurolinguistics, 20(1), 50-64.
4. Scherer, S., Stratou, G., Mahmoud, M., Boberg, J., Gratch, J., Rizzo, A., & Morency, L. P. (2013). Automatic behavior descriptors for psychological disorder analysis. In 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG) (pp. 1-8). IEEE.
5. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.
6. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
7. Zhu, Y., Shang, Y., Shao, Z., & Guo, G. (2017). Automated depression diagnosis based on deep networks to encode facial appearance and dynamics. IEEE Transactions on Affective Computing, 9(4), 578-584.
8. Yang, L., Jiang, D., He, L., Pei, E., Oveneke, M. C., & Sahli, H. (2018). Decision tree based depression classification from audio video and language information. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge (pp. 17-24).
9. Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Breakspear, M., & Parker, G. (2016). Detecting depression: A comparison between spontaneous and read speech. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2749-2753). IEEE.
10. Williamson, J. R., Quatieri, T. F., Helfer, B. S., Ciccarelli, G., & Mehta, D. D. (2019). Vocal and facial biomarkers of depression based on motor incoordination and timing. In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge (pp. 65-72).

11. Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., ... & Pantic, M. (2017). Avec 2017: Real-life depression, and affect recognition workshop and challenge. In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge (pp. 3-9).
12. Du, Z., Li, W., Huang, D., & Wang, Y. (2018). Bipolar disorder recognition via multi-scale discriminative audio temporal representation. In Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop (pp. 23-30).
13. Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10), 1499-1503.
14. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).