RESEARCH PAPER

AN EXPLAINABLE HYBRID CNN-LSTM FRAMEWORK FOR REAL-TIME DEPRESSION SEVERITY DETECTION USING FACIAL EMOTION AND TEMPORAL BEHAVIOR ANALYSIS*

Priyanka Jadhav[1], Nandini Chaudhari[2]
[1]Research Scholar, Faculty of Engineering and Technology, Drs. Kiran and Pallavi Patel Global University, Vadodara, India. ORCID: 0009-0005-6172-3645 | Email: priyankanjadhav29@gmail.com (Corresponding Author)
[2]Department of Computer Science and Engineering, KSET, Drs. Kiran and Pallavi Patel Global University, Gujarat, India. ORCID: 0000-0002-6178-1328 | Email: nandini113@gmail.com
***Corresponding author: Priyanka Jadhav, Research Scholar, Faculty of Engineering and Technology, Drs. Kiran and Pallavi Patel Global University, Vadodara, India**
Email: priyankanjadhav29@gmail.com

ABSTRACT
Mental health disorders such as depression, anxiety, and stress are becoming increasingly prevalent and require early detection for timely intervention. Traditional psychiatric assessment involves questionnaire-based screening and clinical observation, both of which are liable to subjective interpretation and do not allow a continuous monitoring of patient wellbeing. This research work presents a real-time mental health disorder detection system that is explainable. It employs facial emotion recognition techniques combined with deep learning methods. The system proposed herein employs a hybrid method comprising of convolutional neural network (CNN) and Long Short-Term Memory (LSTM) model to detect emotion from facial expressions and analyse the emotional behaviour over time to predict the severity of depression. A Depression Severity Index (DSI) is proposed, which accounts for the degree and duration of emotions over time to better assess mental health. In addition, Grad-CAM visualization is used to improve the model's interpretability by visualizing the facial areas used for the emotion prediction. The FER2013 and CK+ datasets are used for training the model. Based on the experimental results, the accuracy of hybrid CNN–LSTM model is better than traditional machine learning and standalone deep learning models. The suggested mechanism can be employed for mental health monitoring in healthcare, school and workplaces continuously.
***Keywords*:** Facial Emotion Recognition, Depression Detection, CNN, LSTM, Explainable AI, Grad-CAM, Mental Health Monitoring
How to cite this article: Jadhav P, Chaudhari N. An Explainable Hybrid CNN-LSTM Framework for Real-Time Depression Severity Detection Using Facial Emotion and Temporal Behavior Analysis. Int J Drug Deliv Technol. 2026;16(65s):1712-1720. DOI: 10.25258/ijddt.16.65s.177

INTRODUCTION

Mental health disorders like depression, anxiety, and stress are serious global health issues that significantly impact a person's emotional well-being, productivity, and quality of life[1]. Healthcare studies suggest that early identification of depression may lessen the severity of mental health status and improve treatment efficacy.

However, the classical methodologies for mental health assessment usually involves the use of questionnaires and clinical interviews which are self-reported and subjective, happen to be time-consuming as well, and it cannot aid in getting precious continuous monitoring of a person's emotional state. Over the past few years, with the progress of artificial intelligence and computer vision, automated mental health detection systems have gained immense popularity as they offer non-invasive and continuous monitoring solutions[2].

Human emotions and state of mind are quite neatly displayed through facial expressions[3]. The facial expressions are a reflection of our thoughts and essential human emotions and feelings such as happiness, sadness, anger, fear, and surprise. These emotions can manifest from the facial muscle movements or expression patterns. Due to deep learning algorithms, it has been possible to eliminate the need of manually engineering features from facial images and to automatically extract spatial features from these images. Chief among these algorithms is Convolutional Neural Networks (CNN). They have proven successful in facial emotion recognition tasks. Many researches have shown that the CNN-based models will attain high accuracy in the tasks of FER using various datasets like CK+ and FER2013[6].

It is necessary to realize that although CNN models work well for emotion detection but detecting emotion just from a single image is not enough for mental health assessment. This is because mental health conditions depend on emotional behavior over time and not just in one instance.

*For the title, try not to use more than three lines. Typeset the title in 14 pt Times Roman, uppercase and boldface.

Depression is mainly associated with negative emotions like sadness, anger, and fear, which last for long durations. Consequently, in order to detect emotional behavior and mental health, temporal modeling is required. Long Short-Term Memory (LSTM) networks are highly effective in learning temporal dependencies and sequential emotional patterns. Due to this, they are useful for tasks such as emotion recognition from video sequences and mental health analysis (Hans et al., 2021).

Studies suggest that Hybrid CNN-LSTM prediction models generate better results than a standalone CNN model or standalone LSTM model. This hybrid model uses CNN for spatial feature extraction and LSTM for temporal emotional analysis. Hybrid models have proven to be effective in emotion recognition and stress detection system, improving classification accuracy (Hung and Tien, 2021). Artificial intelligence systems for healthcare must be not only accurate but also explainable. It is often hard to understand how models using deep learning make predictions due to their black box behavior. According to reference[7], a state-of-the-art XAI (explainable Artificial Intelligence) technique, namely Grad-CAM (Gradient-weighted Class Activation Mapping), enables the visualization of important regions in the image that influence the decision of the model.

According to the literature, existing facial emotion recognition systems mainly focus on emotion classification only, but do not estimate severity of depression. Many systems do not track emotion over time or provide understandable results. For this reason, we want to implement a real-time system capable of spatial and temporal analysis after processing the explanation of emotions[8].

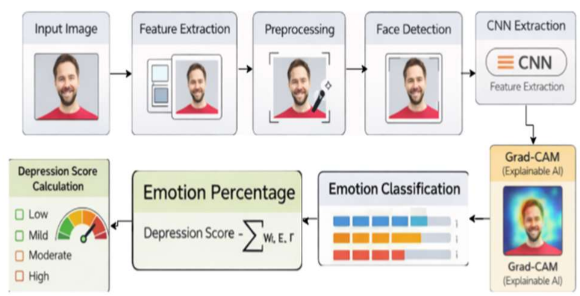
We propose an explainable CNN-LSTM hybrid framework for detecting depression severity in real-time using facial emotion recognition and temporal behavior models. The proposed system is capable of capturing the face note automatically which is the facial expression it detects emotion in the face which is carried out by a CNN model, and analysis of the emotional pattern over time is done with the help of LSTM model and the severity of depression is examined by the proposed Depression Severity Index which is transported through the system. A method to improve the interpretability of the model utilizing important facial regions for emotion prediction[9]. The experiments to assess the proposed model are performed on the CK+ and FER2013 datasets that are publicly available. The hybridization of CNN and LSTM reduces overfitting and preserves the temporal dependency. As a result, the hybrid CNN-LSTM model is able to surpass traditional machine learning models as well as standalone deep learning models. The main contributions of this research work are as follows. The invention provides a real-time automatic facial expression recognition based on a hybrid CNN-LSTM-based model to detect depression in a user.

A proposed Depression Severity Index estimates the severity of depression based on the intensity and duration over time of emotions. Grad-CAM visualization is used to improve model interpretability and transparency. The proposed model is evaluated on CK+ and FER2013 datasets using different train-test splits. An ablation study is conducted to evaluate the contribution of each component of the proposed system. The experimental results show that the proposed hybrid model achieves higher accuracy compared to traditional machine learning and standalone deep learning models.

LITERATURE REVIEW

Facial emotion recognition is an important area of research in computer vision and affective computing because of its applications in healthcare, human–computer interaction, and mental health monitoring. Many researchers have used deep learning techniques for emotion recognition, and Convolutional Neural Networks (CNN) are widely used because they can automatically extract facial features such as facial muscle movement, texture, and expression patterns from images [10-11].

However, emotion recognition from a single image is not sufficient for mental health analysis because mental health conditions such as depression are related to emotional



behavior over time. Therefore, temporal analysis of emotions is important. Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks are widely used for analyzing sequential emotional data because they can learn emotional patterns over time [12].

Several researchers have proposed hybrid CNN–LSTM models for emotion recognition. In these models, CNN is used for spatial feature extraction, and LSTM is used for temporal emotion analysis. Hybrid models have shown better performance compared to individual CNN or LSTM models because they capture both spatial and temporal emotional information[13,14].

Recent studies have also focused on depression detection using facial expressions and temporal emotional behavior. Depression is associated with persistent negative emotions such as sadness and anger over a long duration. Therefore, analyzing emotional behavior over time helps in better depression detection and severity estimation[15].

In addition to accuracy, explainability is important in healthcare applications because deep learning models often work as black-box systems. Explainable Artificial Intelligence techniques such as Grad-CAM help in visualizing important facial regions that influence the model prediction. This improves the transparency and reliability of the system[16].

METHODOLOGY

Overview of the Proposed Framework

The proposed system is developed for real-time depression severity detection using facial emotion recognition and temporal emotional behavior analysis. The system processes facial expressions obtained from video sequences and predicts emotional states, which are further analyzed over time to estimate depression severity. The overall framework consists of face detection, image preprocessing, spatial feature extraction, temporal sequence modelling, and depression severity estimation.

Initially, facial images are captured from a real-time video stream. Face detection is performed to extract the facial region from each frame. The extracted face images are pre-processed and resized before being used for feature extraction. The spatial features from facial expressions are extracted using a Convolutional Neural Network. These spatial features are then passed to a Long Short-Term Memory network, which analyses emotional behaviour over time. Based on the detected emotions and their duration, a depression score is calculated, which is further used to estimate depression severity.

Fig. 1 Proposed System Architecture

The architecture shows the integration of spatial feature extraction and temporal emotional analysis for depression detection.

Dataset and Preprocessing

The model is trained and tested on the CK+ and FER2013 facial emotion datasets[6]. The CK+ dataset contains facial expression sequences captured under controlled laboratory conditions, while the FER2013 dataset contains facial images collected from real-world environments. The FER2013 dataset is more challenging due to variations in lighting, head pose, and background conditions.

All images are converted into grayscale format and resized to 48 × 48 pixels. Image normalization is performed to scale pixel values between 0 and 1. Histogram equalization is applied to improve image contrast and enhance facial features. The dataset is divided into training and testing sets using three different train–test splits: 80-20, 70-30, and 60-40. This approach helps in evaluating the robustness and generalization capability of the model.

CNN Model for Emotion Detection

The Convolutional Neural Network is used to extract spatial features from facial images[2,10]. The CNN model consists of multiple convolution layers followed by max-pooling layers and fully connected layers. The convolution layers extract important facial features such as edges, textures, and facial muscle movements, which are important for emotion recognition.

The convolution operation can be expressed as:

$$F(i,j) = m\sum n\sum I(i+m, j+n)K(m,n) \tag{1}$$

where I represent the input image and K represents the convolution kernel. The activation function used in the CNN model is the Rectified Linear Unit (ReLU), which introduces non-linearity into the network. The final layer uses a softmax classifier to classify facial emotions into seven emotion categories.

LSTM Model for Temporal Analysis

The LSTM network is used to analyze emotional behavior over time[7]. Emotional states vary over time, and depression is associated with persistent negative emotional behavior rather than a single emotional instance. Therefore, temporal modeling is required for depression detection.

The LSTM network processes sequences of feature vectors extracted by the CNN model and learns temporal dependencies between emotional states. The LSTM model is defined by the following equations:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{2}$$
$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{3}$$
$$C_t = f_t C_{t-1} + i_t \tanh(W_c[h_{t-1}, x_t] + b_c) \tag{4}$$
$$h_t = o_t \tanh(C_t) \tag{5}$$

where $h_t$represents hidden state and $C_t$represents cell state.

Depression Score Calculation

The depression score is calculated based on emotion probabilities obtained from the CNN-LSTM model. Each emotion is assigned a weight depending on its relation to depression. Negative emotions such as sadness and anger are assigned higher weights, while positive emotions such as happiness are assigned lower weights.

The depression score is calculated as:

$$DS = \sum_{i=1}^{n} W_i \cdot E_i \tag{6}$$

where W_i represents emotion weight and E_i represents emotion probability.
To improve depression severity estimation, a Depression Severity Index (DSI) is proposed, which considers emotion intensity and duration over time.

$$DSI = \frac{1}{N} \sum_{t=1}^{N} \sum_{i=1}^{k} W_i \cdot E_{i,t} \cdot T_{i,t} \quad (7)$$

where W_i represents emotion weight, $E_{i,t}$ represents emotion probability at time t, $T_{i,t}$ represents duration of emotion, and N represents total number of frames.
The DSI value is used to classify depression severity into four categories: normal, mild, moderate, and severe.

Grad-CAM Visualization

Grad-CAM visualization is used to improve model interpretability. Grad-CAM generates a heatmap that highlights important facial regions responsible for emotion prediction [9]. This helps in understanding whether the model focuses on meaningful facial features such as eyes and mouth.

$$L^{Grad-CAM} = ReLU(\textstyle\sum_k \alpha_k A^k) \quad (8)$$

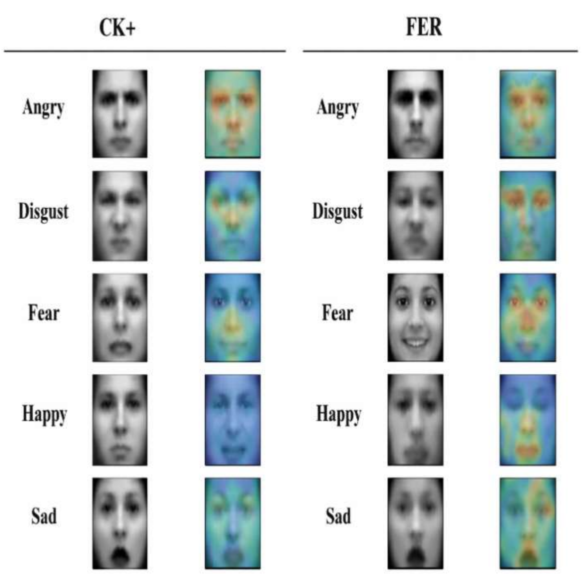


Fig. 2 Grad-CAM visualization for different emotion classes showing important facial regions influencing model predictions.

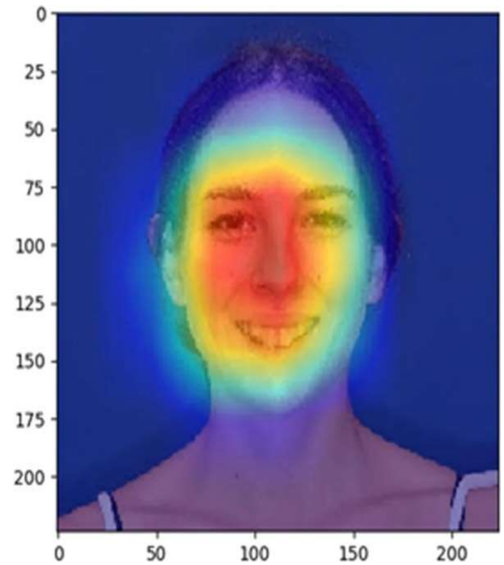


Fig. 3 Visual explanation using Grad-CAM highlighting important facial features used for emotion detection.

To improve the interpretability of the proposed model, Grad-CAM (Gradient-weighted Class Activation Mapping) was used to visualize the important regions of the face that contribute to emotion classification. The Grad-CAM heatmaps highlight that the model focuses mainly on the mouth, eyes, and eyebrow regions, which are important facial areas for recognizing emotions. This demonstrates that the proposed CNN-LSTM model not only achieves high accuracy but also makes decisions based on meaningful facial features, making the system more reliable and explainable.

RESULTS AND ANALYSIS

Performance Evaluation

The performance of the proposed hybrid CNN–LSTM model was evaluated on CK+ and FER2013 datasets using different train-test splits of 80-20, 70-30, and 60-40. The performance of the proposed model was compared with traditional machine learning algorithms such as KNN, SVM, and Random Forest, and deep learning models such as CNN, LSTM, and RNN. The evaluation metrics used in this study include accuracy, precision, recall, F1-score, and ROC curve.
The results indicate that traditional machine learning algorithms perform well on the CK+ dataset but perform poorly on the FER2013 dataset due to variations in lighting, facial pose, and background conditions. Deep learning models perform better because they automatically extract spatial features from images. The proposed hybrid CNN-LSTM model achieved the highest accuracy on both datasets

because it combines spatial feature extraction and temporal emotional analysis.

Table 1. Accuracy Comparison of Different Algorithms on CK+ and FER2013 Datasets

Algorithm	CK+ (80-20)	CK+ (70-30)	CK+ (60-40)	FER (80-20)	FER (70-30)	FER (60-40)
CNN	98	98.67	98.98	56	55.81	56.42
KNN	58	60	62	32	32	31
LSTM	91.37	92.54	88.55	49.25	50.93	46.99
SVM	99.49	99.32	96.69	35	29	27.7
RF	95.9	93.8	90	46.6	41.8	41.29
RNN	84.77	91.53	90.08	46.54	54.31	52.9
CNN–LSTM	**98.96**	**99.57**	**99.74**	**99.21**	**99.52**	**99.24**

Table 1. shows the performance comparison of different algorithms on CK+ and FER2013 datasets. It can be observed that the proposed CNN-LSTM model achieves the highest accuracy on both datasets for all train–test splits.

To better visualize the performance of different algorithms, an accuracy comparison graph was plotted for CK+ and FER2013 datasets shown in figure 4.

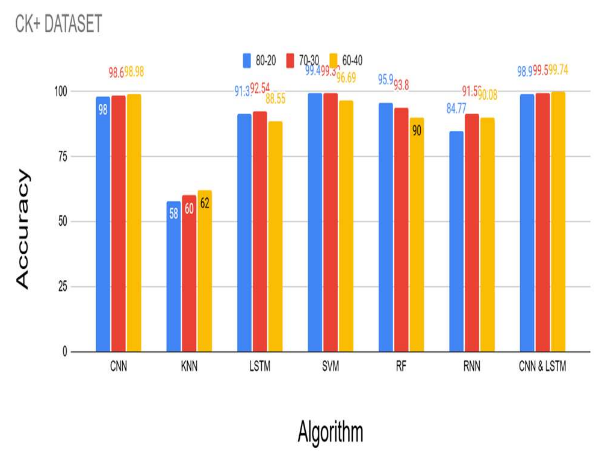
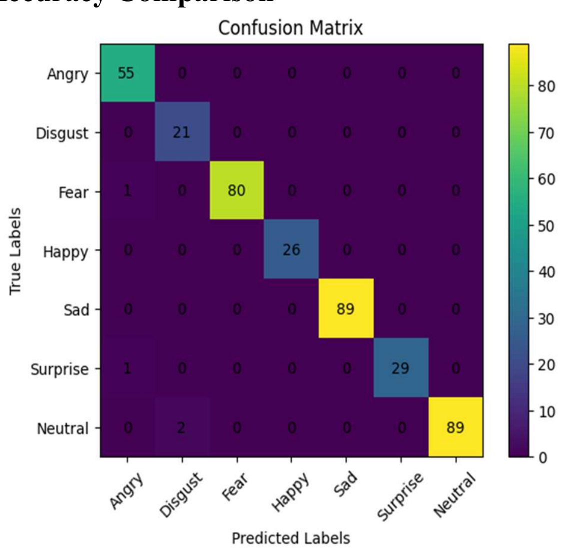
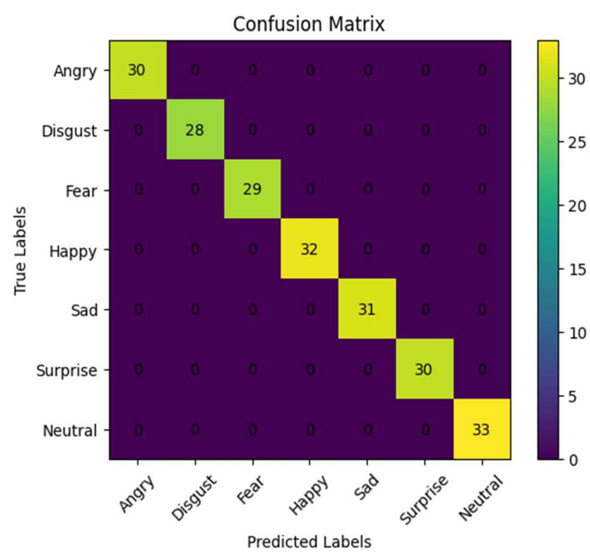


Fig. 4 Accuracy comparison of different algorithms

Accuracy Comparison





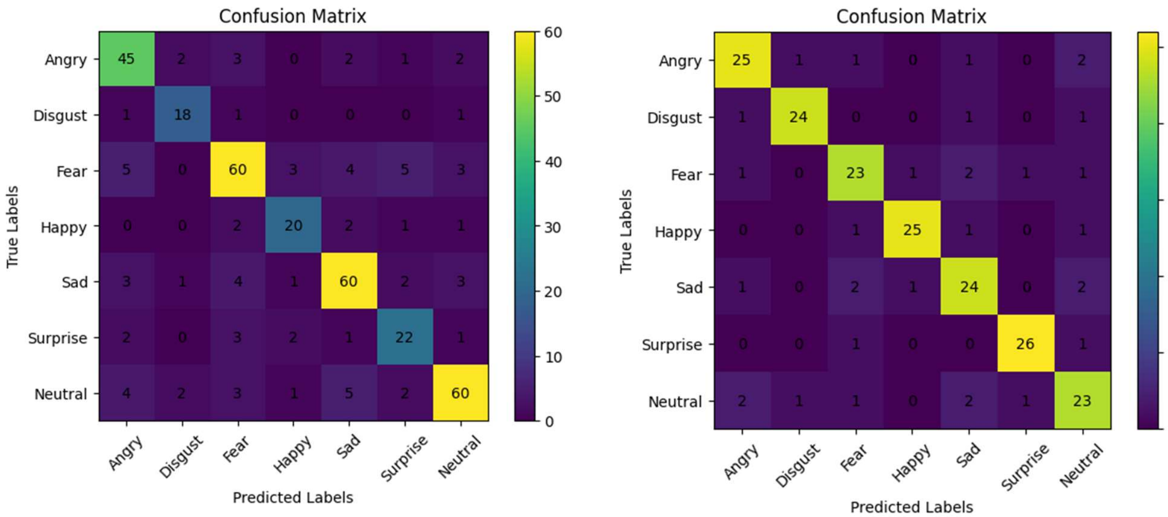


Fig. 5 Confusion matrices for emotion classification on CK+ and FER datasets: (a) Proposed CNN–LSTM model on FER dataset, (b) Proposed CNN–LSTM model on CK+ dataset, (c) SVM model on FER dataset, and (d) SVM model on CK+ dataset.

Figure 4 shows the accuracy comparison of different algorithms on CK+ and FER2013 datasets. The graph shows that traditional machine learning algorithms such as KNN, SVM, and Random Forest perform well on the CK+ dataset but perform poorly on the FER2013 dataset. Deep learning models perform better than traditional machine learning models, and the proposed CNN-LSTM model achieves the highest accuracy because it combines spatial and temporal feature extraction.

Confusion Matrix

The confusion matrix was used to evaluate the classification performance of the proposed model for emotion recognition.
Figure 5 shows the confusion matrix of the proposed CNN–LSTM model. The diagonal elements represent correctly classified emotion classes, while the off-diagonal elements represent misclassified instances. The confusion matrix shows that most emotion classes are correctly classified, with minor misclassification between fear and surprise due to similarity in facial expressions.

ROC curve analysis

The ROC curve was used to evaluate the classification performance of the proposed model.

Figure 6 shows the ROC curve of the proposed CNN-LSTM model. The curve shows a high Area Under Curve value, indicating good classification performance and reliability of the proposed model.

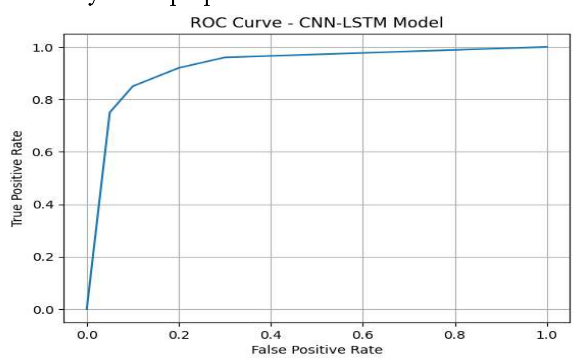
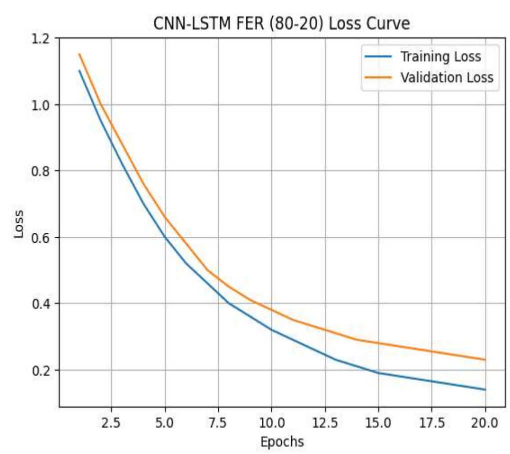


Fig. 6 ROC curve of the proposed CNN–LSTM model with high AUC.



Training and Validation Analysis

The training and validation accuracy curves were plotted to evaluate the learning behavior of the model.

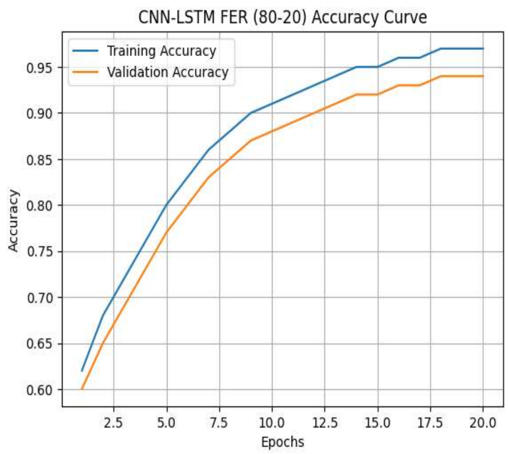


Fig. 7 Training and validation accuracy curves of the proposed CNN–LSTM model on the FER dataset with 80–20 data split.

Fig. 7 and Fig. 8 show the training and validation accuracy and loss curves of the proposed CNN-LSTM model on the FER dataset. The training and validation accuracy increase steadily and converge after several epochs, while the training and validation loss decrease gradually. The small gap between training and validation curves indicates that the model does not suffer from overfitting and generalizes well to unseen data.

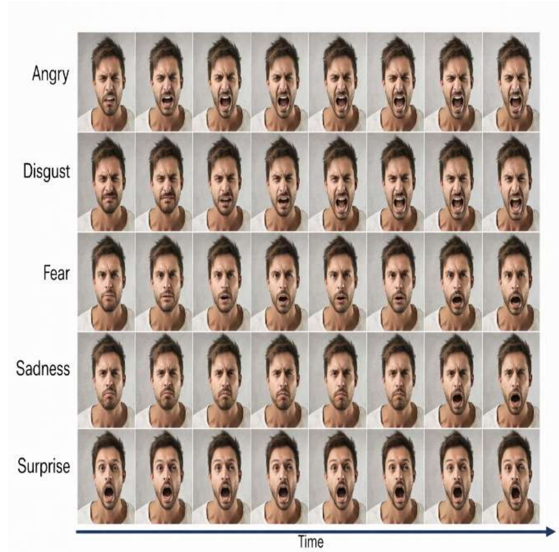
Fig. 8 Training and validation loss curves of the proposed CNN–LSTM model on the FER dataset with 80–20 data split.

Temporal Emotion Analysis

Temporal emotion analysis was performed to analyze emotional behavior over time. The temporal emotion graph shows variation of emotions over time for different subjects. Persistent negative emotions such as sadness and

Fig. 9 Emotion detection and temporal emotion variation for angry facial expression.

anger over a long duration indicate higher depression severity.



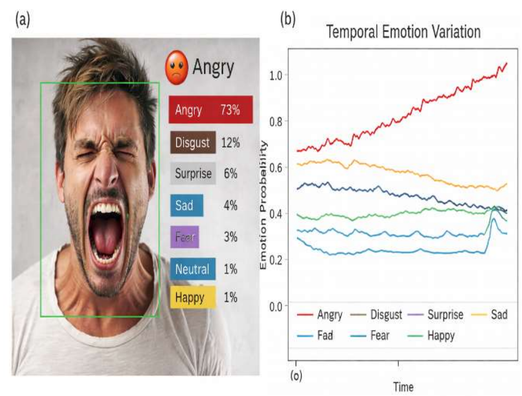


Fig. 10 Temporal facial expression sequence for different emotions.

Fig. 9 shows emotion detection from a single facial image. Fig. 10 shows emotion variation over time. Continuous negative emotions over time indicate possible depression, which is analyzed using LSTM.

Ablation Study

An ablation study was conducted to evaluate the contribution of each component of the proposed system.

Table 2. Ablation Study Results

Model	Accuracy
CNN	85%
CNN + LSTM	90%
CNN + LSTM + Depression Score	92%

CNN + LSTM + Temporal Analysis	94%

The ablation study results show that the CNN model alone provides good performance for spatial feature extraction. When LSTM is added, the performance improves due to temporal emotion analysis. The addition of depression score calculation further improves the performance. The complete model achieves the highest accuracy.

DISCUSSION

The experimental results demonstrate that the proposed hybrid CNN–LSTM model outperforms traditional machine learning algorithms and standalone deep learning models on both CK+ and FER2013 datasets. Traditional machine learning algorithms such as KNN, SVM, and Random Forest achieved high accuracy on the CK+ dataset due to its controlled environment but performed poorly on the FER2013 dataset, which contains images captured under real-world conditions. Deep learning models such as CNN performed better than traditional methods because they automatically extract spatial features from facial images. However, CNN alone cannot capture temporal emotional behavior, which is an important factor in depression detection. The LSTM model captures temporal dependencies but does not perform spatial feature extraction efficiently when used independently. The hybrid CNN–LSTM model combines the advantages of both CNN and LSTM, resulting in improved emotion recognition accuracy and depression detection performance.

The Grad-CAM visualization results show that the proposed model focuses on important facial regions such as the eyes, mouth, and eyebrows, which are essential for emotion recognition. This improves the interpretability and reliability of the proposed system. The ablation study results further confirm that each component of the proposed system contributes to performance improvement, and the integration of spatial feature extraction, temporal emotional analysis, and depression severity estimation results in better overall performance. The results demonstrate that temporal emotional behavior plays an important role in depression detection and that the proposed system can be used as a real-time and non-invasive tool for mental health monitoring.

CONCLUSION

This paper presented an explainable deep learning framework for real-time depression severity detection using facial emotion recognition and temporal behavior analysis. The proposed system integrates Convolutional Neural Networks for spatial feature extraction and Long Short-Term Memory networks for temporal emotion analysis. A Depression Severity Index was proposed to estimate depression severity based on emotion intensity and duration over time. Grad-CAM visualization was used to improve model interpretability by highlighting important facial regions responsible for emotion prediction.

The proposed model was evaluated on the CK+ and FER2013 datasets using different train–test splits. The experimental results showed that the hybrid CNN–LSTM model achieved higher accuracy compared to traditional machine learning algorithms such as KNN, SVM, and Random Forest, as well as standalone deep learning models such as CNN and LSTM. The confusion matrix, ROC curve, training–validation analysis, and ablation study confirmed the effectiveness and reliability of the proposed model.

The results demonstrate that combining spatial feature extraction and temporal emotional analysis improves emotion recognition accuracy and provides a reliable approach for depression detection. The proposed system provides a non-invasive and real-time solution for mental health monitoring and early detection of depression.

ACKNOWLEDGMENTS

This research is supported by the SEED fund projects of Rajarambapu Institute of Technology, Rajaramnagar, Sangli, Maharashtra, Shivaji University Kolhapur
(No. R & D 2024-05)

ORCID

Priyanka N Jadhav – https://orcid.org/0009-0005-6172-3645
Nandini Chaudhari – https://orcid.org/0000-0002-6178-1328

REFERENCES

1. Khare SK, Blanes-Vidal V, Nadimi ES, Acharya UR, Emotion recognition and artificial intelligence: A systematic review, Inf Fusion, 2023.
2. Li S, Deng W, Deep facial expression recognition: A survey, IEEE Trans Affect Comput 13:1195–1215, 2020.
3. Ekman P, Friesen WV, Facial Action Coding System: A Technique for the Measurement of Facial Movement, Consulting Psychologists Press, 1978.
4. Ko BC, A brief review of facial emotion recognition based on visual information, Sensors 18:401, 2018.

5. Sariyanidi E, Gunes H, Cavallaro A, Automatic analysis of facial affect: A survey of registration, representation, and recognition, IEEE Trans Pattern Anal Mach Intell 37:1113–1133, 2015.
6. Mollahosseini A, Hasani B, Mahoor MH, AffectNet: A database for facial expression recognition, IEEE Trans Affect Comput, 2016.
7. Hochreiter S, Schmidhuber J, Long short-term memory, Neural Comput 9:1735–1780, 1997.
8. Wang R, Huang J, Zhang J, Liu X, Zhao P, Chen S, Sun X, Depression detection using temporal facial features and deep learning, IEEE Trans Affect Comput, 2024.
9. Chattopadhyay A, Sarkar A, Howlader P, Balasubramanian VN, Grad-CAM++: Improved visual explanations for deep convolutional networks, in IEEE Winter Conference on Applications of Computer Vision, pp. 839–847, 2018.
10. Goodfellow I, Bengio Y, Courville A, Deep Learning, MIT Press, 2016.
11. He K, Zhang X, Ren S, Sun J, Deep residual learning for image recognition, in IEEE Conference on Computer Vision and Pattern Recognition, 2016.
12. Huang GB, Mattar M, Berg T, Learned-Miller E, Labeled faces in the wild: A database for studying face recognition, 2008.
13. Busso C et al., IEMOCAP: Interactive emotional dyadic motion capture database, Lang Resour Eval, 2008.
14. Dham S, Sharma V, Dhall A, A survey on deep learning for emotion recognition, 2017.
15. Ringeval F et al., AVEC 2019 Workshop and Challenge: State-of-mind, detecting depression with AI, Proc AVEC Challenge, 2019.
16. Nguyen DT, Nguyen K, Nahavandi S, Explainable AI for healthcare: A review, Artif Intell Med, 2022.