<ISCCSC 2025>
<Conference Series > <Volume number and Year> <DOI Number>

Attention Enhanced Hybrid CNN – LSTM Framework for Realtime Mental Health Monitoring and Support

N. C. Brintha, M. Harishkanth, M. Susikumar, R. Yabin, S. Finix

Department of Information Technology,

Kalasalingam Academy of Research and Education,

Krishnankovil, Tamil Nadu, India

Brinthachris2k@gmail.com, harishkanth166@gmail.com,

susikumar231@gmail.com, yabin0111@gmail.com, finixthalmetha@gmail.com

Abstract.

Mental health disorders such as depression are among the most pressure challenges in healthcare today. Traditional methods of diagnosis rely very much on the volunteering questionnaire and subjective assessment by the physician, which often gives rise to underdiagnosis or delayed intervention. To address this challenge, we propose a medical therapy assist system that integrates Artificial Intelligence (AI) and Convolutional Neural Network techniques to automatically detect and assess the level of depression in patients. The system takes advantage of the Convolutional Neural Network (CNN) to remove facial emotions from live video stream and Long Short-Term Memory (LSTM) with attention to analyze temporal emotion trends. The solution further covers a web-based platform where patients can participate in live emotion monitoring, self-assessment questionnaire, and can get personal treatment, while physicians can see analytics, therapy history and generate detailed reports. Our proposed system crosses the limits of existing static assessment by enabling the monitoring of real -time, continuous and intelligent depression. It provides an efficient framework, which ensures a high level of data privacy, providing a efficient structure to increase 96.8% accuracy and reliability in health care systems, ensuring trust and reliability within the quality of medical data handling and care

Keywords. Artificial Intelligence, CNN, LSTM, Attention Mechanism, Depression Severity Assessment, Facial Emotion Recognition, Mental Health Monitoring, Medical Therapy Assist System.

1. INTRODUCTION

These rapid increases in the incidences of depression, anxiety, and other related psychological issues have brought forth the need for a supportive solution that helps both patients as well as medical staff in effectively detecting and treating these issues.

Depression is currently being experienced by tens of millions of people around the globe, but the social stigma attached to psychological issues tends to create a lag in their effective treatment. Normally, these tests only include interviews with the patient as well as some assessments, which are entirely dependent upon personal perspectives. But with recent breakthroughs in artificial intelligence and machine learning, automated systems for psychological analysis of human behaviors and emotions through certain techniques of analysis have become possible. Deep learning models for the analysis of human expressions through facial emotion recognition techniques have shown significant success in finding psychological states of humans like happiness, sadness, anger, fear, and surprise through facial expressions. But human emotional expressions change over time, so temporal analysis is important for psychological analysis. To overcome this issue, the combination of Convolutional Neural Networks with Long Short-Term Memory helps the system in learning both spatial features of emotion as well as their temporal evolution. This helps in the monitoring of the patient as well as the categorization of the severity of depression into various levels such as normal, mild, moderate, or severe. It helps fill the gap between technology and the medical field by allowing a medical therapy support system for various interactions between the patient as well as the doctor through the use of AI-based analysis.

2. LITERATURE SURVEY

FA.H. Shah et al proposed a framework that addresses a scenario where a user seeks to locate a specific individual using only a visual memory of that person. In this approach, the user selects images of the person based on target-specific information, which are then fed into a trained model for automatic recognition. A hybrid CNN-BiLSTM architecture to recognize facial emotions, offered by Alina Alamichhane and Gopal Karn, was aimed at eliminating the temporary reliance following the extraction of features. Their model integrates both convolutional neural networks to learn spatial features and bidirectional LSTM layers to learn sequential features which are interconnected with those that emphasize the idea of temporal conversion by using LSTM-based processing. Jiamin Liu, Yuehu Liu, and Yuanqi Su proposed a multimodal emotion recognition system, which incorporates LSTM-based recurrent neural networks with attention system. Their process integrates wide range input features and dynamically focuses on most emotionally important temporal patterns enhancing the capacity of the model to make sense of emotion sequence over time.The CNN-based model that was developed by Dan Dan Yan, Lu Lu Jhao, Shin Wang Songs, Jio Han Jung and Lee Cae Yang took advantage of EEG data to identify clinical depression automatically. Their suggested networks, ENENET, Dipkonwnet and Shalocon-weight, exhibited that convolutional architectures could identify the severity of depression through the neural signal. Even though their method is based on the EEG, instead of facial or questionnaire-based data, it successfully shows how CNN is able to predict the level of depressive moods. Yi Li, Zidacai, and Jingyi Wang introduced a CNN–LSTM framework for depression index prediction using clinical data such as the MadRS score. The fusion of convolutional and recurrent layers enables both spatial and temporal feature learning, aligning closely with this study’s aim of integrating emotion outputs with questionnaire data to estimate depression levels categorized as mild, medium, or severe.

3. METHODOLOGY

3.1. Proposed CNN–LSTM with Attention Framework

The proposed system will estimate the severity of depression in real time by employing an attention-enhanced CNN-LSTM architecture. Facial expressions are captured via either a web or mobile camera, and features that relate to the emotions will be extracted from each video frame using a Convolutional Neural Network. These features represent spatial emotional information contained within the facial expressions. The frame-level features are fed into a Long Short-Term Memory network combined with an attention mechanism in a row. This attention module lets the model underline emotionally important time segments and thus capture meaningful temporal variation in emotional behavior. By this selective focus, it enhances the system's ability for recognizing emotional trends associated with the level of depression severity. Based on the temporal patterns learned, the model classified individuals into normal, mild, moderate, and severe depression. If there is no facial input or if the input is unreliable, the system switches over to a previously validated self-assessment questionnaire. This ensures uninterrupted analysis and maintains the reliability of depression assessment across varied usage conditions.

3.2. Dataset and Preprocessing

The model proposed uses a publicly accessible dataset containing facial expressions with emotions labeled for multiple people. All images are first adjusted to the same resolution to ensure consistency during the feature extraction process using convolutional techniques. To increase the model's robustness and reduce overfitting, techniques such as image mirroring, changing contrast, and random rotation are applied to improve data augmentation. Afterward, the dataset is split into three distinct parts, which are used for training the model, validating its performance, and assessing its final effectiveness.

4. FINDINGS AND DISCUSSION

4.1. Quantitative Results Summary

From the experimental results, it is evident that the Attention-Enhanced CNN-LSTM framework outperforms other approaches like CNN, LSTM, and CNN-LSTM without the attention mechanism. As presented in Table 1, the proposed model demonstrates strong performance, achieving an overall accuracy of 96.8% and maintaining high levels of precision, recall, and F1-measure. The consistent improvement across all these metrics indicates that combining spatial feature extraction with temporal learning, supported by an attention mechanism, is effective for accurately predicting the severity levels of depression.

Table 1: Performance Metrics of the proposed CNN–LSTM with attention Model

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC

Proposed CNN LSTM + Attention	96.8	96.2	95.8	95.9	0.982
CNN-Only Baseline	89.4	88.1	87.9	87.8	0.903
LSTM-Only Baseline	83.2	82.5	81.0	81.6	0.872
CNN–LSTM (Without Attention)	92.7	91.5	91.2	91.3	0.941

4.2. Discussion

The findings emphasize the idea that the CNN-based spatial feature learning and LSTM-based temporal sequence modeling have significant positive effects on depression detection. It is also the attention mechanism that increases the reliability of classification through focusing on emotionally informative segments of time. The proposed framework is competitive or better accurate than current methods reported in the literature because it does not use complicated physiological sensors, which makes it applicable to real-time mental health monitoring in an application.

4.3. Strengths and Limitations

The main advantage of the proposed system consists in its capability to learn dynamic emotional patterns with the help of an attention-enhanced CNN-LSTM architecture that leads to a high level of classification and stable results. Nevertheless, the system can also worsen in performance when there is poor lighting or when using a low-resolution camera, and additional large-scale clinical testing is needed to test its ability in the real world.

5. CONCLUSION

The assignment is built upon the development of a medical therapy assist system, which connects computer vision, deep learning, and healthcare processes to complement mental health monitoring. CNN, combined with LSTM, can enable the system to recognize, not only track the progress of individual feelings, which can help in the process of diagnosing depression more accurately. The practical implementation of the device and its user friendliness is provided by the web-based implementation that comprises of the patient and physician dashboard. The level of reliability is also raised by Fallback questionnaire

mechanism and automated report generation system. Finally, the system provides an all-inclusive AI-managed platform to patients and physicians and presents the divide between traditional medical care and the new AI-managed healthcare.

Data Availability Statement

The data sets to be used in this research are facial expression data sets that were found on open-source repositories. There was no proprietary or personally identifiable data in this study.

6. REFERENCES

[1] X. Zhao, X. Liang, and Z. Liu, "LSTM-based emotion recognition using facial expression and physiological signals," *Sensors*, vol. 21, no. 17, p. 5750, 2021.

[2] J. Liu, Y. Liu, and Y. Su, "Multi-modal emotion recognition with temporal-band attention based on LSTM-RNN," in *Intelligent Computing and Internet of Things*, Springer, pp. 168–177, 2018.

[3] D. D. Yan, L. L. Zhao, X. W. Song, X. H. Zang, and L. C. Yang, "Automated detection of clinical depression based on deep convolutional neural networks using EEG signals," *Biomed. Eng./Biomed. Tech.*, vol. 66, no. 5, pp. 443–456, 2021.

[4] H. Zhang, J. Zhang, and B. Song, "Hybrid CNN-LSTM network for speech emotion recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 647–651, 2020.

[5] R. P. Pathak, H. Gangwar, A. Agarwal, and M. L. Kolhe, "Detecting negative emotions to counter depression using CNN," in *Proc. International Conference on Artificial Intelligence Applications*, Springer, pp. 345–354, 2021.

[6] Y. Zhang, H. Wu, and J. Zhao, "Automatic depression detection via facial expression recognition using CNN and transfer learning," *Frontiers in Psychiatry*, vol. 12, p. 665443, 2021.

[7] J. Singh, L. B. Saheer, and O. Faust, "Speech emotion recognition using attention model," *Computers*, vol. 12, no. 4, p. 81, 2023.

[8] S. M. Tiwari and M. S. Alam, "Facial emotion recognition using deep learning," International Journal of Research in Applied Science and Engineering Technology (IJRASET), vol. 9, no. 4, pp. 2182–2186, 2021.

[9] S. M. Khare, M. Goswami, and A. Khare, "Deep learning-based EEG emotion recognition: current trends and future directions," Frontiers in Psychology, vol. 14, p. 1126994, 2023.

[10] Y. Zhang, H. Yuan, and X. Li, "EEG-based emotion recognition using multi-scale dynamic CNN and gated transformer," Scientific Reports, vol. 14, p. 31319, 2024.