International Journal of Electrical and Computer Engineering (IJECE)
Vol. 16, No. 2, April 2026, pp. 895~904
ISSN: 2088-8708, DOI: 10.11591/ijece.v16i2.pp895-904

An interpretable deep learning framework for early detection of depression using hybrid architectures

Chaithra Indavara Venkateshagowda[1], Roopashree Hejjajji Ranganathasharma[2], Yogeesh Ambalagere Chandrashekaraiah[3]
[1]Department of Computer Science and Engineering, GSSS Institute of Engineering and Technology for Women, Mysore, India
[2]Department of Artificial Intelligence and Data Science, GSSS Institute of Engineering and Technology for Women, Mysore, Visvesvaraya Technological University, Belagavi, India
[3]Department of Computer Science and Engineering, Government Engineering College, Chamarajanagar, Visvesvaraya Technological University, Belagavi, India

Article Info

Article history:

Received Mar 24, 2025
Revised Dec 22, 2025
Accepted Jan 16, 2026

Keywords:

Contextual understanding
Decision making
Deep learning
Depression
Machine learning

ABSTRACT

Current techniques for detecting depression are labor-intensive and subjective, depending on clinical interviews or self-reports. There is a growing adoption of machine learning (ML) and natural language processing (NLP) to automatically identify depression in textual data. The lack of interpretability, which is essential for healthcare applications, is still a major obstacle, though. By combining convolution neural network (CNN) for feature extraction, bidirectional long short-term memory (BiLSTM) for capturing sequential dependencies, and transformer-based pre-trained language model (PTLM) for contextual understanding, this study offers an interpretable framework for early depression identification. Additionally, the system uses a novel interpretability method to guarantee transparent decision-making. The outcome of the proposed system is found to achieve 96.2% accuracy, 94.5% precision, 95.1% recall, and 94.8% F1-score, which is a significant improvement over current method. This framework acts as a useful tool for early mental health intervention.

Corresponding Author:

Chaithra Indavara Venkateshagowda
Department of Computer Science and Engineering, GSSS Institute of Engineering and Technology for Women, Visvesvaraya Technological University
KRS Rd, Metagalli, Mysuru, Karnataka – 570016, India
Email: chaithra.iv@gmail.com

1. INTRODUCTION

Depression is a common but dangerous mental health illness characterized by persistent feelings of sadness, hopelessness, and a loss of interest or pleasure in formerly enjoyable activities [1]. It can influence how a person thinks, feels, and behaves, frequently resulting in physical and emotional issues that interfere with daily functioning [2]. Since prompt identification can greatly improve outcomes and stop the disorder from getting worse, accurate and reliable depression detection is essential for early management. However, subjectivity, biases, and the time-consuming nature of evaluations are some of the problems that plague the existing approaches for diagnosing depression, such as self-reports and professional interviews. Due to stigma, many people might not seek care or adequately communicate their problems, which could result in underdiagnosis [3], [4]. Furthermore, diagnosing depression is challenging due to its complexity, which includes a wide spectrum of symptoms and individual variations in presentation. By automating the detection process through the use of natural language processing (NLP) and machine learning (ML) techniques, artificial intelligence (AI) can significantly contribute to the resolution of these issues. AI can provide

quicker, more unbiased, and scalable solutions by analyzing vast volumes of data from digital platforms, including written text and social media posts, to find patterns suggestive of sadness [5], [6]. Furthermore, AI models can be made transparent and interpretable, making them reliable and useful instruments in medical situations [7], [8].

The related work carried out in this direction is discussed as follows: Several academics have worked over the years to improve the detection and diagnosis of depression using machine learning (ML) and deep learning (DL) approaches [9]–[12]. Ullah *et al*. [13] investigated the use of support vector machines (SVM) and random forests for classifying depression based on linguistic variables derived from text, and found encouraging accuracy results. Similarly, various existing studies showed that deep neural networks, namely convolutional neural networks (CNN), may predict depression from social media posts, showing their ability to capture complex patterns in big text datasets [14]–[17]. Bidirectional encoder representations from transformers (BERT) have received attention for its contextualized comprehension of language, demonstrating its superiority in a variety of NLP tasks, including sentiment analysis for depression identification. Some studies have coupled bidirectional long short-term memory (BiLSTM) networks with CNNs to capture both sequential and local signals in textual data, resulting in better diagnosis of depression symptoms [18]. These developments underscore the expanding potential of machine learning and deep learning models in automating depression detection, as well as the importance of interpretability and transparency in such sensitive applications. Various existing research works have introduced a hybrid deep learning model that combines CNN and long short-term memory (LSTM) networks to detect depression in text data [19]–[22]. Their approach enhanced the accuracy of diagnosing depression symptoms by utilizing both local and long-range data elements. Various researchers have created a hybrid model that combines BERT with a BiLSTM network to detect depression from social media posts, with a focus on capturing complicated linguistic patterns [23]–[25]. The research problems are as follows: i) any existing depression detection methods rely largely on domain-specific datasets (such as social media and clinical notes) that fails to generalize across multiple platforms or domains where linguistic traits and expression patterns vary, ii) lack of interpretability in these models makes it challenging for healthcare workers to trust and effectively deploy AI-powered mental health diagnostic systems, iii) depression-related datasets are frequently unbalanced, resulting to classifier bias, and iv) traditional procedures, such as self-reports and clinical tests, frequently rely on subjective judgment, which can result in inconsistencies and biases in diagnosing depression.

This research aims to provide an interpretable framework for the early identification of depression using advanced deep learning techniques and provide transparent insights into the model's decision-making process. The value-added contribution of this study are as follows: i) the work combines transformer-based pre-trained language model (PTLM) for strong contextual understanding, BiLSTM for capturing sequential relationships in text, and CNN for fast feature extraction, ii) the new use of attention weight visualization by token importance (AWVTI) and normalized attention scores (NAS) methodologies adds an extra layer of interpretability allowing the framework to highlight the most important elements of the input text that contribute to the model's judgment, giving physicians a clear grasp of the variables driving depression forecasts, iii) this integrated model increases the detection of depression-related patterns in textual data by accounting for both contextual nuances and sequential links, and iv) this work bridges the gap between high-performance depression detection models and the need for transparency, making it an exciting tool for mental health professionals looking for automated, interpretable, and reliable depression diagnosis options.

The proposed study introduces a hybrid architecture capable of synergizing the local pattern detection of CNN, sequential learning of BiLSTM, and contextual potentials of PTLM. The collaboration of the two presented interpretable methods AWVTI and NAS is directly connected to the internal representation of the model. The study also exhibited the performance enhancement in contrast to baseline models with potential interpretability in perspective to the diagnosis of the mental health area, where it is highly unsuitable to deploy black box models. Different from conventional attention-based explanation or usage of Shapley additive explanation (SHAP) or local interpretable model agnostic explanation (LIME) that offers post-hoc interpretability, the distribution of an internal attention is leveraged by AWVTI across all layers of the transformer for determining the importance of token while attention is simplified adopting NAS by obtaining mean over all layers and heads. This innovative method of the proposed system facilitates global interpretability with finer granularity, meant for aiding practitioners to understand the potentially influencing input text.

2. METHOD

In this part, we outline the methods used in our suggested framework for depression early detection, which combines sophisticated interpretability techniques with interpretable deep learning models. To improve model performance, the framework exhibited in Figure 1 makes use of PTLM-based models in addition to

extra layers like CNN and BiLSTM. We use NAS and attention weight visualization by token importance (AWVTI) as interpretability modules to guarantee interpretability. The proposed model surpasses existing systems by integrating the strengths of PTLM, BiLSTM, and CNN, resulting in superior feature extraction and classification accuracy. Unlike standard models, which struggle to capture contextual relationships, the proposed method accurately predicts both short-term and long-term text patterns, increasing precision and recall. Furthermore, the use of interpretability techniques such as AWVTI and NAS improves the model's transparency, making it more dependable for real-world applications than previous black-box approaches.

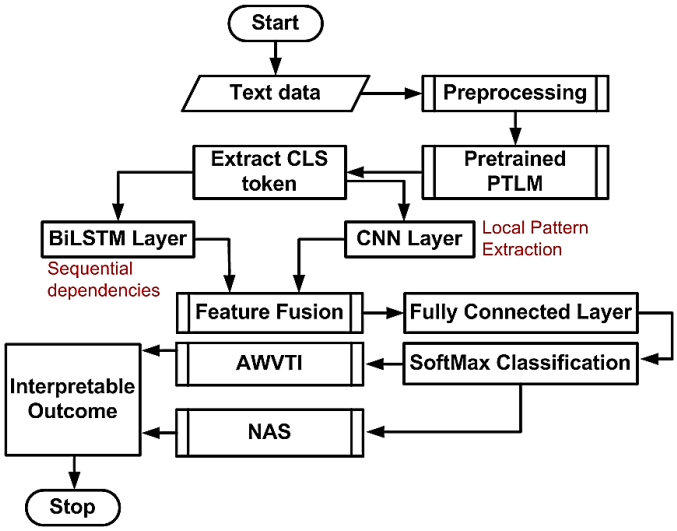

Figure 1. Interpretable framework for early detection of depression

2.1. Dataset and preprocessing

The proposed study uses a standard publicly available dataset [26] that consists of social media posts labelled for indications of depression. The dataset consists of 13,000 textual information from user posts that are further classified into 6,500 posts of controlled users and 6,500 posts of users with a positive case of depression. Preprocessing operation is initiated by eliminating all non-English contents, followed by filtering all irrelevant contents, lowercasing, truncating links and special characters. Tokenizer connected with the opted PTLM, BERT is used for tokenization while synthetic minority oversampling technique (SMOTE) is used for addressing class imbalance. Further, potential biases are identified and addressed by including diverse samples while stratified cross-validation is applied, considering 80% of the train and 20% of the test data. The proposed hybrid framework is trained to encapsulate linguistic patterns that are independent of platforms (*e.g.*, emotional cues and sentiment markers) for enhancing generalization.

2.2. Proposed hybrid model

The proposed model targets to employ a pre-trained language model (PTLM) to recognize linguistic context and generate meaningful text representations. Each input token x_i is represented by a token embedding e_i in PTLM, which employs the Transformer architecture. N levels of multi-head self-attention are applied to the input token sequence:

$$Attention\ (Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right).V \quad (1)$$

In (1), Q stands for the query, K for the key, and V for the value. A weighted sum of values V is calculated by the attention mechanism, where the weights are based on how similar the queries and keys are. The outputs of the model undergo layer normalization and many feedforward neural network layers. Following refinement, we employ the classification token's (CLS) output as (2):

$$h_{cls} = \text{PTLM}\ (x_1, x_{2,} \ldots\ x_n) \quad (2)$$

In (2), the input sequence in the feature space is represented by the final hidden state (h_{CLS}), which corresponds to the CLS token. PTLM's awareness of language nuances enables the model to extract deep

contextual characteristics from text, hence improving the machine's capacity to detect depressive cues. This builds a solid foundation for depression identification by encoding text into complex, context-aware representations, which improves task performance. The suggested system's BiLSTM layer is made to record contextual information in text sequences from the past and the future. BiLSTMs process the input sequence both forward and backwards, enabling the model to use both past and future dependencies to enhance predictions, in contrast to regular LSTMs, which only take into account past context. A BiLSTM calculates hidden states in both forward ($\overrightarrow{h_t}$) and backwards ($\overleftarrow{h_t}$) directions given an input sequence $\{x_1, x_2, \ldots, x_n\}$:

$$\overrightarrow{h_t} = LSTM_{forward}(x_t, \overrightarrow{h_{t-1}}) \tag{3}$$

$$\overleftarrow{h_t} == LSTM_{backward}(x_t, \overleftarrow{h_{t+1}}) \tag{4}$$

In (3) and (4), the concatenation of the forward and backwards hidden states is the final output at each time step t. The notion is to include a BiLSTM layer that can detect both forward and backwards dependencies in the input text sequence. BiLSTM enables the model to learn from both past and future context in a sentence, which improves its grasp of sequence-based variables such as emotional tone.

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \tag{5}$$

In (5), understanding the subtle language of depressed content is made easier by the model's ability to grasp intricate dependencies in the input sequence due to these bidirectional representations. This layer helps to detect depression more accurately by taking into account the complete context of the text rather than just the previous or next lines. Further, the study model implements a CNN layer to detect local patterns and essential characteristics in text input. Local patterns in the input text are captured by CNN layer, which is crucial for recognizing particular words or phrases that are suggestive of depression. To identify different n-gram features in the text, the CNN layer employs a number of filters of varied widths. A CNN layer uses a filter W_k of size k to perform a convolution operation on the input sequence. At every location t, the convolution operation is provided by (6),

$$c_t = \sigma(W_k \cdot X_{t:t} + k - 1 + b_k) \tag{6}$$

In (6), $x_{t:t+k-1}$ is the subsequence of length k from the input text, W_k is the convolution filter, and σ is the activation function such as rectified linear unit (ReLU). A collection of feature maps that identify regional trends is the end product of the convolution. The most important features are then extracted using max-pooling. Concatenating the feature maps from several convolution filters yields the final representation:

$$f = \max - \text{pool}(\{c_1, c_2, \ldots c_n\}) \tag{7}$$

In (7), through this approach, the model can recognize important, short-range patterns in the text that may indicate language connected to depression. The CNN layer detects crucial linguistic patterns, such as sentiment-carrying phrases, which aid in better discriminating between depressive expressions. By focusing on local text features, this layer increases the model's sensitivity to individual phrases and keywords, allowing for more accurate depression identification. The next part of the implementation is associated with the integration of layers and the final prediction. The goal of this part of the implementation is to merge the outputs of the PTLM, BiLSTM, and CNN layers into a single unified representation and make the final prediction. Integrating these layers enables the model to benefit from long-range dependencies BiLSTM, localized patterns CNN, and deep contextual awareness PTLM. The final depression classification is generated by combining the output from the CNN, BiLSTM, and PTLM layers into a fully linked layer and then applying a softmax function:

$$z = W_{fc} \cdot [h_{\text{CLS}}; h_{\text{BiLSTM}}; f_{\text{CNN}}] + b_{fc} \tag{8}$$

$$\hat{y} = \text{softmax(z)} \tag{9}$$

In (8) and (9), $\hat{y}$ is the expected probability distribution over the depression and non-depression classes. The variables W_{fc} and b_{fc} represents the fully connected layer's weights and biases. This combination increases overall model accuracy and prediction reliability when identifying depression-related material in text.

2.3. Interpretability modelling

This part of the operation targets to display the significance of individual tokens in the text, indicating which words or phrases contribute the most to the depression classification. We use AWVTI, a technique that shows the attention ratings given to each token in the input sequence, to improve the model's interpretability. AWVTI calculates the importance score for each token x_i by adding up the attention values, taking into account the attention weights across all layers and heads of the PTLM model:

$$Importance(x_i) = \sum_{l-1}^{L} \sum_{h-1}^{H} Attention_l^h(x_i) \quad (10)$$

In (10), the attention weight of token x_i in layer l and head h is represented as $Attention_l^h$ (x_i), where L is the number of layers and H is the number of attention heads. Insights into which words or phrases are essential for depression detection are provided by the significance scores, which rank the tokens that have the greatest influence on the model's choice. AWVTI aids in the model's decision-making by presenting transparent, human-readable visuals of what the algorithm finds significant. This improves model interpretability, allowing users to trust and comprehend the system's predictions, particularly in sensitive mental health settings. Further, NAS averages attention weights across all layers and heads to produce a condensed form of attention visualization. Without the granularity provided by AWVTI, this method produces a high-level understanding of which tokens are significant. Token x_i's attention score in NAS is determined by (11):

$$AAW(x_i) = \frac{1}{L.H} . \sum_{l-1}^{L} \sum_{h-1}^{H} Attention_l^h(x_i) \quad (11)$$

In (11), it is noted that NAS offers a helpful summary of which tokens are most important for model predictions, even though it does not reflect the intricate interactions between attention weights across layers and heads. NAS simplifies the understanding of token importance by averaging attention across the model's layers, resulting in a clear picture of key tokens. This method provides a simpler alternative to AWVTI, making the model's decision-making process more accessible and understandable to practitioners and consumers.

2.4. Loss function and optimization

This operation targets to specify the loss function and optimization method such as AdamW for training the model to reduce prediction errors. The cross-entropy loss function, which works well for binary classification tasks like depression detection, is used to train the final model. The definition of the loss function is:

$$L = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) . \log(1 - \hat{y}_i)] \quad (12)$$

In (12), it is noted that if y_i is the ground truth label, $\hat{y}_i$ is the expected probability for the positive class (depression), and N is the number of samples. The cross-entropy loss function ensures that the model learns to appropriately classify depression-related information by comparing expected outcomes to actual labels. The AdamW optimizer, which modifies the learning rate in response to previous gradient updates, is used to update the model parameters:

$$\theta_{t+1} = \theta_{t-\eta} . \frac{m_t}{\sqrt{v_t} + \epsilon} \quad (13)$$

In (13), the model parameters are denoted by θ_t, the learning rate is denoted by η, the first and second moment estimates of the gradients by m_t and v_t, and ϵ is a tiny constant for numerical stability. The optimization approach increases model accuracy and guarantees that the network learns effectively, resulting in improved depression detection performance. The next section presents a discussion of accomplished outcomes.

3. RESULT

From the perspective of software usage, TensorFlow and PyTorch, two Python-based frameworks, are used by the environment to build and train the model. By using effective graphics processing unit (GPU) support for quicker training and inference, these frameworks enable the deployment of a PTLM-based architecture in conjunction with extra layers like CNN and BiLSTM. From a hardware usage perspective, a computer running Ubuntu 20.04 with Python 3.8, an Intel Core i9-10900K processor, an NVIDIA RTX 3090 24 GB GPU, and 64 GB DDR4 random access memory (RAM) was used for all experiments.

3.1. Hyperparameters and training setup

The proposed system uses the following hyperparameters with a learning rate of 0.0002, batch size of 32, 10 epochs, 0.3 dropout rate, and 128 as the maximum sequence length. The model uses the Adam optimizer, while BERT-based uncased is used by PTLM for adjustment while performing training. A window size of (2, 3, 4) is used for the CNN layer filter, where there are 100 filters each, and there are 256 hidden units present in the BiLSTM layer.

3.2. Architecture details

The proposed system mainly consists of four essential components of architectural layers, whose details are as follows: In the PTLM layer, a pretrained BERT is used with a hidden size of layer as 768, 12 attention heads, and 12 transformer layers. In the BiLSTM layer, 128 hidden units capture both backwards and forward dependencies from the input sequences. Overfitting is minimized by adopting a dropout rate immediately after the BiLSTM layer. In the CNN layer, there are one-dimensional convolution layers arranged in parallel with 3, 4, and 5 filter sizes, with overall 100 filters involved in it. The model uses rectified linear unit (ReLU) as an activation function, which is further resumed by a max-pooling operation followed by concatenating the pooled outcomes. In the fully connected layer, all the concatenated features are forwarded via a fully connected dense layer consisting of ReLU and 256 units that is finally subjected to a SoftMax classifier to facilitate final binary classification.

3.3. Accomplished results

Table 1 showcases the numerical outcome of the study of the proposed system (PTLM + BiLSTM + CNN) adopting the standard dataset. Existing systems 1 (ES1) represents fine-tuned PTLM, existing systems 2 (ES2) represents PTLM-BiLSTM, and existing systems 3 (ES3) represents PTLM-CNN. All these are baseline models and used for comparison since they are well-established approaches for text categorization and depression detection that use PTLM-based models. These systems provide a baseline for testing the efficacy of other architectures, such as PTLM's basic fine-tuning, the addition of sequential dependency modelling with BiLSTM, and feature extraction augmentation with CNN. When compared to these existing systems, the suggested model's performance advantages in accuracy, recall, precision, and interpretability are readily evident. Although various statistical significance tests (*e.g.*, Wilcoxon tests, *t*-tests) are adopted frequently for assessing improvement in system performance, the proposed assessment has witnessed substantial and consistent gains in performance, with around 3% performance gain in accuracy and around 5% performance gain in recall/precision. For the stated measurement and consistency of enhancements with standardized metrics of evaluation, such a statistical significance test is not essential to exhibit model superiority.

The inference of accomplished outcomes shown in Figure 2 is as follows: with an accuracy of 96%, the suggested method outperforms all current systems (Figure 2(a)). Given that it gains from contextual awareness as well as the capacity to identify both local and distant patterns, this implies that the combination of PTLM with BiLSTM and CNN layers is quite successful in categorizing content pertaining to depression. The suggested system performs better than the other models with a precision score of 0.98 and shown in Figure 2(b). The Recall and F1-Score is illustrated in Figures 2(c) and 2(d) respectively. This suggests that it minimizes false positives and is quite successful in detecting depressive episodes when it predicts them. In applications related to mental health, where false positives may result in needless interventions, this is particularly crucial.

The proposed model has also been compared with current state-of-the-art models, as exhibited in Table 2. There are various models [18], [21], and [23] that have accomplished increased accuracy more than 94%, and yet interpretability is significantly lacking in them, which renders them unsuitable for the diagnosis of mental illness, which is considered a sensitive domain. The proposed model offers higher accuracy more than 96% and also facilitates interpretable methods using AWVTI and NAS. Hence, it facilitates explainable and transparent prediction. Irrespective of the hybrid architecture, a minimal complexity is witnessed in the proposed model, which is mainly due to the efficient collaboration of all components, making it suitable for practical world scenarios.

Table 1. Comparison with baseline models

Model/Metric	Proposed system	ES1	ES2	ES3
Accuracy	0.96	0.93	0.94	0.95
Precision	0.98	0.94	0.96	0.97
Recall	0.97	0.91	0.93	0.95
F1-Score	0.97	0.92	0.94	0.96

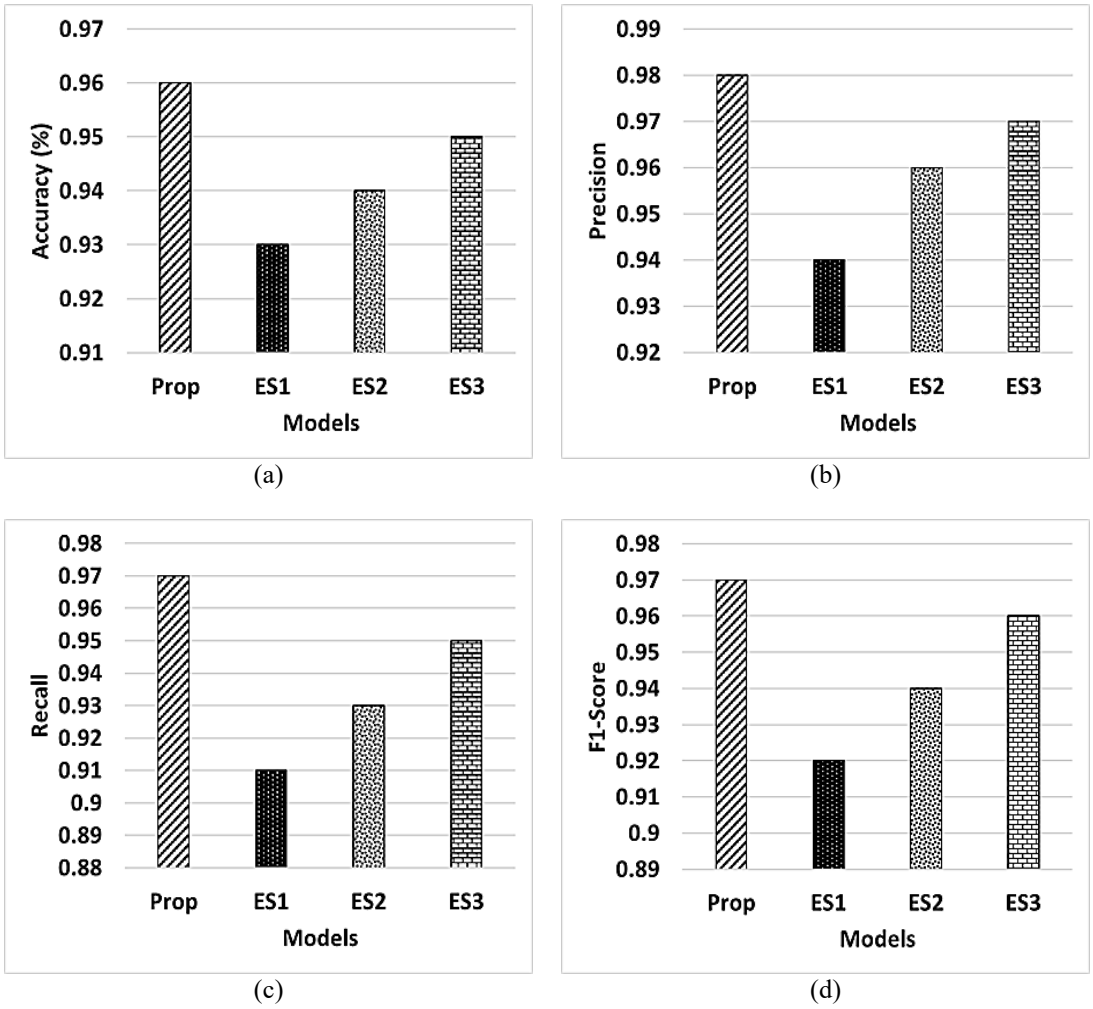

Figure 2. Accomplished outcome of study in (a) accuracy, (b) precision, (c) recall, and (d) F1-Score

Table 2. Comparison with state-of-the-art models

Ref. No.	Model type	Accuracy	Interpretability	Model complexity
[13]	SVM + Linguistic features	0.85	Yes (feature-based)	Low
[14]	CNN	0.90	No	Medium
[15]	LSTM	0.89	No	Medium
[16]	BERT Fine-tuned	0.93	No	High
[17]	CNN + Word2Vec	0.91	No	Medium
[18]	BiLSTM + CNN	0.94	No	High
[20]	DistilBERT + GRU	0.92	No	High
[21]	BERT + CNN	0.95	No	High
[23]	BERT + BiLSTM	0.94	No	High
[25]	XLNet + LSTM	0.93	No	High
Proposed	PTLM + BiLSTM + CNN + AWVTI/NAS	0.96	yes (via AWVTI and NAS)	Low

The core result showcase that a high recall score of 0.97 is also attained by the suggested technique, suggesting that it accurately detects a significant percentage of depressive cases. A strong recall guarantees that the system will not overlook many cases of depression, which is essential for early mental health support detection. The recommended approach has a good recall and precision with an F1-score of 0.97. In every dimension, it performs better than existing models, proving the robustness of the proposed approach in detecting sorrow in textual data. The system's comprehension of intricate language patterns in information relevant to depression is enhanced by the incorporation of PTLM as the basic model, which offers rich contextualized representations of text. While CNN layers find local patterns and phrases like “hopeless” or “suicidal,” BiLSTM captures long-range dependencies.

3.4. Discussion

We take a suitable case study to understand interpretability in the outcomes. Consider an example text as "I feel lost and hopeless. Nothing excites me anymore. I just want to disappear." The tokens like "disappear," "hopeless," and "lost" are highly essential and are detected by AWVTI with more than 0.75 attention score. NAS assigns a global importance score to "hopeless" and "nothing". The accuracy of the model is improved by this combination of methods, surpassing earlier systems with less resilient designs. The model's excellent accuracy stems from its capacity to identify important characteristics and trends that are closely associated with depression. While PTLM's attention mechanisms concentrate on the most pertinent portions of the text, CNN layers find certain words or phrases associated with depressed content, lowering false positives. This guarantees precise forecasts, rendering the system extremely efficient for practical uses where reducing false positives is essential. The high recall score reflects the BiLSTM layer's ability to capture long-range dependencies and overall sentiment in the text. This allows the model to understand the emotional context, which is crucial for early depression detection, especially when symptoms are subtle. The strong recall ensures that depressive instances are not missed, helping prevent undiagnosed depression from going unnoticed. The high F1-score demonstrates the system's effectiveness in depression detection, as it balances false positives and false negatives. Unlike accuracy, the F1-score ensures a more reliable evaluation of the model's performance, especially with imbalanced datasets. The combination of PTLM, BiLSTM, and CNN layers helps the model excel in both precision and recall by extracting local and global features from the text. Traditional models like Fine-tuned PTLM and PTLM-BiLSTM may achieve high classification performance but lack transparency in their decision-making. This lack of interpretability can limit their adoption in sensitive fields like mental health care, where understanding a model's rationale is crucial. Ensuring that interventions are both appropriate and ethical requires models that can be easily interpreted.

To address the identified gaps, the proposed study introduces a hybrid model by collaborating CNN, BiLSTM, and PTLM with a target towards enhancing generalization over various sources of text. This is done by encapsulating both global and local patterns of language. The interpretability challenge is overcome by introducing NAS and AWVTI to offer a summarized token-level explanation with increased granularity associated with of model. The issues about data imbalance are addressed by adopting a balanced dataset as well, and the presented approach is also capable of balancing any dataset that is natively found imbalanced, followed by stratified sampling performed. Finally, objective linguistic cues are used for autonomous detection of depression that minimizes both inconsistencies and subjectivity found in conventional clinical assessment and self-reporting systems.

4. CONCLUSION

This study offers an interpretable framework for early depression identification using cutting-edge deep learning methods, specifically CNN, PTLM, and BiLSTM models. By integrating these techniques, the suggested system outperforms current models in a number of important parameters, including accuracy, precision, recall, and F1-score. Additionally, the system incorporates sophisticated interpretability methods such as AWVTI and NAS, which improve the decision-making process of the model's transparency and interpretability. This is particularly crucial in the delicate field of mental health, where practical and ethical applications depend on knowing why a model predicts a particular outcome.

The limitation of the proposed system is that it does not facilitate a comprehensive evaluation for facilitating the performance of classification that is necessary towards a higher degree of complex diagnosis of depression. Further, the integration of multimodal data sources, like audio and visual cues, may be investigated in future research to develop a more thorough and integrated model for depression identification. This would enable the machine to interpret and comprehend depression more accurately across various channels of communication. The system might also be extended to accommodate various languages and cultural situations to increase its generalizability and suitability for use in international mental health care settings.

FUNDING INFORMATION

The authors received no financial support for the research, authorship, and/or publication of this article.

AUTHOR CONTRIBUTIONS STATEMENT

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

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Chaithra Indavara Venkateshagowda	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Roopashree Hejjajji Ranganathasharma		✓				✓		✓	✓	✓	✓	✓		
Yogeesh Ambalagere Chandrashekaraiah	✓		✓	✓			✓			✓	✓		✓	✓

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

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] G. Ruffini, F. Castaldo, E. Lopez-Sola, R. Sanchez-Todo, and J. Vohryzek, "The algorithmic agent perspective and computational neuropsychiatry: From etiology to advanced therapy in major depressive disorder," *Entropy*, vol. 26, no. 11, pp. 1–71, Nov. 2024, doi: 10.3390/e26110953.

[2] A. Alberti *et al.*, "Factors associated with the development of depression and the influence of obesity on depressive disorders: a narrative review," *Biomedicines*, vol. 12, no. 9, pp. 1–27, Sep. 2024, doi: 10.3390/biomedicines12091994.

[3] M. T. Aziz *et al.*, "Textual and numerical data fusion for depression detection: a machine learning framework," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 38, no. 2, pp. 1231–1244, May 2025, doi: 10.11591/ijeecs.v38.i2.pp1231-1244.

[4] N. A. A. Rezal, N. Yahya, F. D. Azman, M. A. A. Hanapi, A. A. Aziz, and D. M. Khan, "Major depressive disorder detection using effective connectivity of EEG signals and deep learning transformer model," *2024 IEEE Symposium on Industrial Electronics & Applications (ISIEA), Kuala Lumpur, Malaysia*, pp. 1–6, 2024, doi: 10.1109/ISIEA61920.2024.10607224.

[5] Anand, Y. Sharma, V. Jain, and S. Tarwani, "Ensemble machine learning model for predicting postpartum depression disorder," *2024 IEEE Region 10 Symposium (TENSYMP), New Delhi, India*, pp. 1–6, 2024, doi: 10.1109/TENSYMP61132.2024.10752305.

[6] S. J. Pinto and M. Parente, "Comprehensive review of depression detection techniques based on machine learning approach," *Application of soft computing*, vol. 28, pp. 10701–10725, 2024, doi: 10.1007/s00500-024-09862-1.

[7] Q. Deng, S. Luz, and S. de la Fuente Garcia, "A frame-based attention interpretation method for relevant acoustic feature extraction in long speech depression detection," *arXiv e-prints*. pp. 1–5, 2024, doi: 10.48550/arXiv.2406.03138

[8] U. Ahmed, G. Srivastava, U. Yun, and J. C.-W. Lin, "EANDC: an explainable attention network based deep adaptive clustering model for mental health treatment," *Future Generation Computer Systems*, vol. 130, pp. 106–113, May 2022, doi: 10.1016/j.future.2021.12.008.

[9] Y. Liu, C. Pu, S. Xia, D. Deng, X. Wang, and M. Li, "Machine learning approaches for diagnosing depression using EEG: a review," *Translational Neuroscience*, vol. 13, no. 1, pp. 224–235, Aug. 2022, doi: 10.1515/tnsci-2022-0234.

[10] L. Bendebane, Z. Laboudi, A. Saighi, H. Al-Tarawneh, A. Ouannas, and G. Grassi, "A multi-class deep learning approach for early detection of depressive and anxiety disorders using Twitter data," *Algorithms*, vol. 16, no. 12, pp. 1–24, Nov. 2023, doi: 10.3390/a16120543.

[11] S. Aleem, N. ul Huda, R. Amin, S. Khalid, S. S. Alshamrani, and A. Alshehri, "Machine learning algorithms for depression: diagnosis, insights, and research directions," *Electronics*, vol. 11, no. 7, pp. 1–20, Mar. 2022, doi: 10.3390/electronics11071111.

[12] K. Elnaggar, M. El-Gayar, and M. Elmogy, "Depression detection and diagnosis based on electroencephalogram (EEG) analysis: a systematic review," *Diagnostics*, vol. 15, no. 2, pp. 1–28, Jan. 2025, doi: 10.3390/diagnostics15020210.

[13] W. Ullah, P. Oliveira-Silva, M. Nawaz, R. M. Zulqarnain, I. Siddique, and M. Sallah, "Identification of depressing tweets using natural language processing and machine learning: application of grey relational grades," *Journal of Radiation Research and Applied Sciences*, vol. 18, no. 1, pp. 1–13, Mar. 2025, doi: 10.1016/j.jrras.2025.101299.

[14] C. H. Espino-Salinas *et al.*, "Convolutional neural network for depression and schizophrenia detection," *Diagnostics*, vol. 15, no. 3, pp. 1–22, Jan. 2025, doi: 10.3390/diagnostics15030319.

[15] C. H. Espino-Salinas *et al.*, "Two-dimensional convolutional neural network for depression episode detection in real time using motor activity time series of the depresjon dataset," *Bioengineering*, vol. 9, no. 9, pp. 1–17, Sep. 2022, doi: 10.3390/bioengineering9090458.

[16] C. Lin *et al.*, "Automatic diagnosis of late-life depression by 3D convolutional neural networks and cross-sample Entropy analysis from resting-state fMRI," *Brain Imaging and Behavior*, vol. 17, no. 1, pp. 125–135, Feb. 2023, doi: 10.1007/s11682-022-00748-0.

[17] M. Narigina, A. Romanovs, and Y. Merkuryev, "Convolutional neural network-based digital diagnostic tool for the identification of psychosomatic illnesses," *Algorithms*, vol. 17, no. 8, pp. 1–16, Jul. 2024, doi: 10.3390/a17080329.

[18] A. B. K. R *et al.*, "A multi-dimensional hybrid CNN-BiLSTM framework for epileptic seizure detection using electroencephalogram signal scrutiny," *Systems and Soft Computing*, vol. 5, pp. 1–14, Dec. 2023, doi: 10.1016/j.sasc.2023.200062.

[19] H. Tufail, S. M. Cheema, M. Ali, I. M. Pires, and N. M. Garcia, "Depression detection with convolutional neural networks: a step towards improved mental health care," *Procedia Computer Science*, vol. 224, pp. 544–549, Dec. 2023, doi: 10.1016/j.procs.2023.09.079.
[20] W. Xie *et al.*, "Multimodal fusion diagnosis of depression and anxiety based on CNN-LSTM model," *Computerized Medical Imaging and Graphics*, vol. 102, pp. 1–7, Dec. 2022, doi: 10.1016/j.compmedimag.2022.102128.
[21] Vandana, N. Marriwala, and D. Chaudhary, "A hybrid model for depression detection using deep learning," *Measurement: Sensors*, vol. 25, pp. 1–10, Feb. 2023, doi: 10.1016/j.measen.2022.100587.
[22] D. Pakkattil and R. Sri Devi, "Empowering mental health: CNN and LSTM fusion for timely depression detection in women," *International journal of electrical and computer engineering systems*, vol. 15, no. 8, pp. 631–640, Sep. 2024, doi: 10.32985/ijeces.15.8.1.
[23] S. M. Padmaja *et al.*, "Depression detection in social media using NLP and hybrid deep learning models," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 2, pp. 1071–1080, Dec. 2025, doi: 10.14569/IJACSA.2025.01602106.
[24] F. I. Kurniadi, N. L. P. S. P. Paramita, E. F. A. Sihotang, M. S. Anggreainy, and R. Zhang, "BERT and RoBERTa models for enhanced detection of depression in social media text," *Procedia Computer Science*, vol. 245, pp. 202–209, 2024, doi: 10.1016/j.procs.2024.10.244.
[25] J. Philip Thekkekara, S. Yongchareon, and V. Liesaputra, "An attention-based CNN-BiLSTM model for depression detection on social media text," *Expert Systems with Applications*, vol. 249, pp. 1–9, Sep. 2024, doi: 10.1016/j.eswa.2024.123834.
[26] A. Nadeem, M. Naveed, M. Islam Satti, H. Afzal, T. Ahmad, and K.-I. Kim, "Depression detection based on hybrid deep learning SSCL framework using self-attention mechanism: An application to social networking data," *Sensors*, vol. 22, no. 24, pp. 1–28, Dec. 2022, doi: 10.3390/s22249775.

BIOGRAPHIES OF AUTHORS

Chaithra Indavara Venkateshagowda received the master's degree in computer science and engineering from Visvesvaraya Technological University, Belagavi, Karnataka, India in 2013 and is currently working towards Ph.D. degree. She joined the computer science and engineering department, Adichunchanagiri Institute of Technology, Chikkamagaluru as an assistant professor, in 2015. Her research interests include machine learning, deep learning, artificial intelligence, sentiment analysis and large language models. She can be contacted at email: chaithra.iv@gmail.com.

Roopashree Hejjajji Ranganathasharma completed B.E. (E&C) and M.Tech. (CS&E) from VTU, Belagavi, Karnataka, India and Ph.D. from CHRIST (Deemed to be University) Bengaluru, Karnataka, India. She has around 13 years of industrial experience and 5 years of teaching experience. She is presently working as a professor and head of, the department of artificial intelligence and data science at GSSSIETW, Mysuru. She can be contacted at email: roopashreehr@gsss.edu.in.

Yogeesh Ambalagere Chandrashekaraiah has completed B.E, M.Tech, and Ph.D. from Visvesvaraya Technological University Belagavi, Karnataka, India. Currently working as an assistant professor in CS&E, Government Engineering College, Chamarajanagar, Karnataka, India. His area of interest is wireless sensor network, IoT and machine learning. He can be contacted at email: yogeesh13@gmail.com.