A Review on Machine Learning and Deep Learning Models for Identifying Potential Depression on Social Media Text Datasets

Sonali Parmar[1], Dr. Sanmati Jain[2]
Department of CSE[1,2]
VITM, Indore, India[1,2]

Abstract: Depression is a global mental health crisis, affecting over 300 million people worldwide, according to the World Health Organization (WHO). Despite its prevalence, depression often goes undiagnosed or is misdiagnosed due to social stigma, limited access to healthcare, and the subjective nature of traditional diagnostic methods. This highlights the urgent need for innovative and objective tools to identify depression early and accurately. Machine learning (ML), with its ability to process and analyze vast amounts of data, offers a promising solution to bridge this gap and transform depression detection methods. Since the data sets are often extremely large and complex, hence off late, opinion mining and sentiment analysis is executed based on artificial intelligence and machine learning models, rather than statistical rule based systems. This paper presents a systematic review of the existing methods for depression detection along with their salient contribution.

Keywords: Depression Detection, Machine Learning, Deep Learning, Tokenization, Semantic Analysis, Classification Accuracy.

I. INTRODUCTION

The integration of Machine Learning (ML) into mental healthcare represents a transformative shift from subjective clinical assessments to objective, data-driven diagnostics. Traditionally, depression is diagnosed through patient self-reports and clinician observations, methods that are often hindered by memory bias or social stigma. Machine learning addresses these limitations by identifying "digital biomarkers"—subtle patterns in speech, text, and physical activity—that are often imperceptible to the human eye. By training algorithms on vast datasets of healthy versus depressed individuals, researchers can now develop screening tools that provide early warnings, potentially reaching individuals who might never seek traditional help. At the core of this technological shift are diverse data modalities, with social media and linguistic analysis being among the most prominent [1].

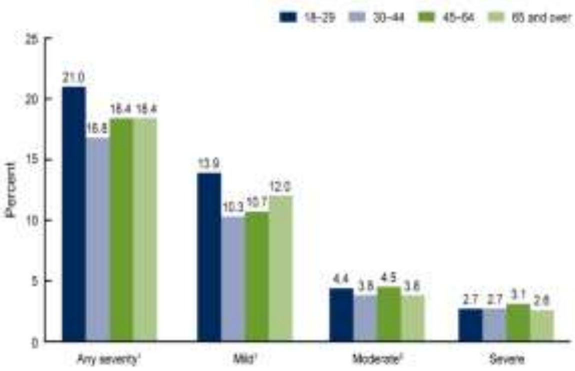

Fig.1.Age-wise onset of depression (Source: https://www.cdc.gov/nchs/products/databriefs/db379.htm)

Severity of depression symptoms was based on the eight-item Patient Health Questionnaire depression scale (PHQ–8), and summarized into none-minimal (values 0–4), mild (values 5–9), moderate (values 10–14), and severe (values 15–24). Those categorized as having no or minimal symptoms of depression are not shown in this figure. Any severity includes those categorized as having either mild, moderate, or severe symptoms of depression in the past 2 weeks.

Severity of depression symptoms was based on the eight-item Patient Health Questionnaire depression scale (PHQ–8), and summarized into none-minimal (values 0–4), mild (values 5–9), moderate (values 10–14), and severe (values 15–24). Those categorized as having no or minimal symptoms of depression are not shown in this figure. Any severity includes those categorized as having either mild, moderate, or severe symptoms of depression in the past 2 weeks [2].

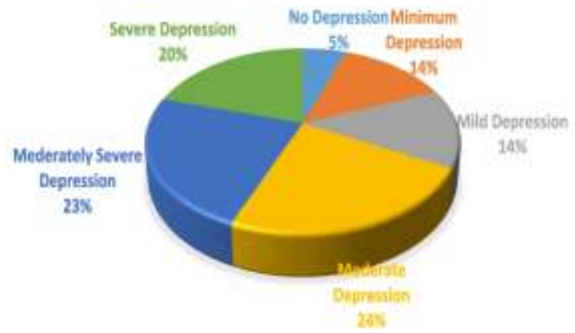

Fig.2 Distribution of depression cases (Source: https://link.springer.com/article/10.1186/s12888-023-05333-3/figures/1)

Displacement is a psychologically stressful event. The displaced individuals may be subject to stressful events such as torture, rape, assassination, and even ethnic cleansing in their home countries and throughout their journey. Displaced survivors who experienced multiple war events perceived multiple negative effects of war on their life domains related to individuals who lived in war areas. Considering the effects, displacement may be accepted as a public health problem and a type of disaster, as it may lead to loss of resources, economic uncertainty, absence of health services and education, insouciant compensation for fundamental human requirements, and disintegration of public structure

Women were more likely than men to experience symptoms of depression across all severity levels. While there was no significant trend by age among adults experiencing severe symptoms of depression, adults aged 18–29 and those aged 65 and over were most likely to experience mild symptoms of depression. A higher percentage of adults aged 45–64 experienced moderate symptoms of depression compared with those aged 30–44 and 65 and over. Adults aged 18–29 were as likely to experience moderate symptoms of depression as those aged 45–64, but the observed differences with the other age groups were not significant. Non-Hispanic white and non-Hispanic black adults were most likely to experience mild symptoms of depression, compared with Hispanic and non-Hispanic Asian adults. Non-Hispanic Asian adults were least likely to experience mild, moderate, or severe symptoms of depression, compared with Hispanic, non-Hispanic white, and non-Hispanic black adults [3].

Based on the eight-item Patient Health Questionnaire depression scale (PHQ–8) and summarized into no or minimal (values 0–4), mild (values 5–9), moderate (values 10–14), and severe (values 15–24) symptoms of depression (2,3). Sample adults were asked how often they have been bothered by the following symptoms in the past 2 weeks: “Little interest or pleasure in doing things;” “Feeling down, depressed, or hopeless;” “Trouble falling or staying asleep, or sleeping too much;” “Feeling tired or having little energy;” “Poor appetite or overeating;” “Feeling bad about yourself, or that you are a failure, or have let yourself or your family down;” “Trouble concentrating on things, such as reading the newspaper or watching television;” and “Moving or speaking so slowly that other people could have noticed? Or the opposite, being so fidgety or restless that you have been moving around a lot more than usual.” Response options were “not at all,” “several days,” “more than half the days,” and “nearly every day,” scored as 0 to 3 points, respectively, and then summed into a total score. Sample adults with two or more PHQ–8 questions answered as “refused,” “don’t know,” or whose answers were not ascertained, were not included in this analysis [4].

II. MACHINE LEARNING AND DEEP LEARNING MODELS FOR IDENTIFICATION OF DEPRESSION

Using Natural Language Processing (NLP), machine learning models analyze the syntax and sentiment of a person's digital footprint. Individuals experiencing depression often exhibit specific linguistic markers, such as an increased use of first-person singular pronouns (e.g., "I," "me") and "absolutist" words like "always" or "completely." Traditional ML models, such as Support Vector Machines (SVM) and Random Forests, are frequently employed to categorize these features, offering a balance between high predictive accuracy and the interpretability required for clinical trust [5].

While classical machine learning relies on manual feature engineering, Deep Learning (DL) has elevated detection capabilities by processing raw, unstructured data. Advanced architectures like Transformers (e.g., BERT and RoBERTa) allow for a sophisticated understanding of context, enabling the AI to distinguish between clinical despair and temporary sadness in text. Similarly, Convolutional Neural

Networks (CNNs) are utilized to analyze facial expressions and vocal spectrograms, detecting "flat affect" or reduced pitch variation—common physiological indicators of depression. These deep learning techniques thrive on complexity, uncovering non-linear relationships in data that traditional statistics might overlook.
This section presents the fundamental machine learning models for identifying depression on social media text data [6].

Support Vector Machine (SVM):
Before the advent of deep learning, traditional machine learning models such as Support Vector Machines (SVM), Decision Trees, Random Forests, and K-Nearest Neighbors (KNN) were widely used for fault detection. The SVM classifies based on the hyperplane.

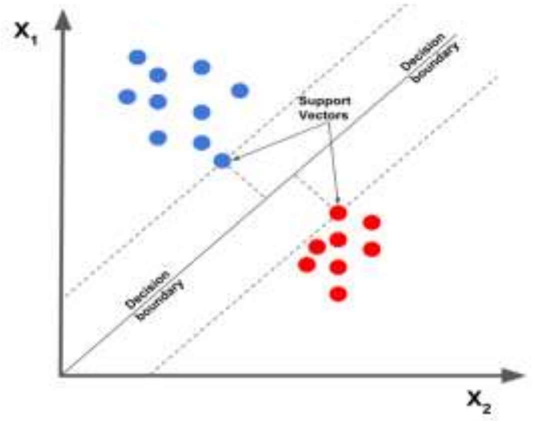

Fig.3 The SVM Model

Figure 3 depicts the SVM Model.

The large variation in the liner scale is replaced with the exponential kernel as [7]:

$$K(X,X') = e^{-\gamma|X-X'|^2} \quad (1)$$

Here,
γ is called the free parameter of RBF
σ is called the feature factor
K represents the RBF Kernel
X *and* X' are the samples in an input feature space
$|X-X'|$ is termed as the Euclidean Distance

The selection of the hyperplane H is done on the basis of the maximum value or separation in the Euclidean distance d given by:

$$d = \sqrt{x_1^2 + \cdots \ldots \ldots . x_n^2} \quad (2)$$

Here,
x represents the separation of a sample space variables or features of the data vector,
n is the total number of such variables
d is the Euclidean distance

The (n-1) dimensional hyperplane classifies the data into categories based on the maximum separation. For a classification into one of **'m'** categories, the hyperplane lies at the maximum separation of the data vector **'X'.** The categorization of a new sample **'z'** is done based on the inequality [8]:

$$d_x^z = Min(d_{C1}^z, d_{C2}^z \ldots d_{C2=m}^z) \quad (3)$$

Here,
d_x^z is the minimum separation of a new data sample from **'m'** separate categories
$d_{C1}^z, d_{C2}^z \ldots d_{C2=m}^z$ are the Euclidean distances of the new data sample **'z'** from m separate data categories.

For instance, SVMs are effective for binary classification tasks, such as distinguishing between urban and rural areas, while Random Forests are used for multi-class classification problems, such as land cover mapping. However, these models struggle with complex patterns in high-resolution imagery and require extensive feature engineering, which limits their scalability and accuracy

ARIMA:
In an autoregressive integrated moving average model commonly known as the ARIMA model assumes that the future value of a variable can be linearly modelled as a function previous samples of the variables and errors of prediction.

$$y_t = \theta_0 + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \varphi_p y_{t-p} + \cdots \ldots \ldots \ldots . \theta_q \varepsilon_{t-q} \quad (4)$$

Here,
y_t is the value of the output variable at time 't'
ε is the prediction error
θ *and* φ are called the model parameters
p *and* q are called the orders of the model

One of ARIMA's key strengths lies in its ability to handle both stationary and non-stationary data. While the ARIMA model assumes the input time series is stationary (i.e., its statistical properties like mean and variance remain constant over time), it incorporates differencing techniques to convert non-stationary data into a stationary format. This makes it highly adaptable for real-world datasets that often exhibit trends or seasonality.

Neural Networks:

Owing to the need of non-linearity in the separation of data classes, one of the most powerful classifiers which have become popular is the artificial neural network (ANN). The neural networks are capable to implement non-linear classification along with steep learning rates. The neural network tries to emulate the human brain's functioning based on the fact that it can process parallel data streams and can learn and adapt as the data changes. This is done through the updates in the weights and activation functions.

The neural network is a connection of such artificial neurons which are connected or stacked with each other as layers. The neural networks can be used for both regression and classification problems based on the type of data that is fed to them. Typically the neural networks have 3 major conceptual layers which are the input layer, hidden layer and output layer.

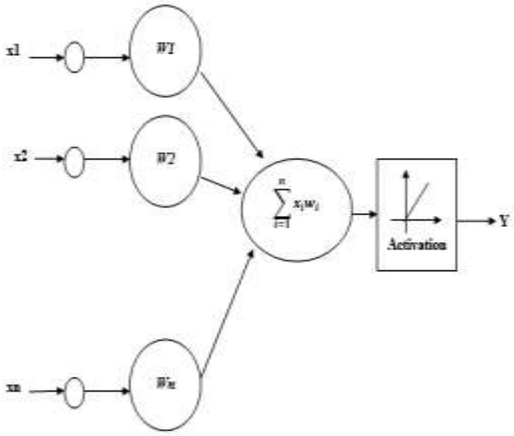

Fig.4 The ANN Model

Figure 4 depicts the ANN model.
The input-output relation of a CNN is given by:

$$y = f\left(\sum_{i=1}^{n} x_i w_i + b\right) \tag{5}$$

Here,
x denote the parallel inputs
y represents the output
w represents the bias
f represents the activation function

The parallel inputs are fed to the input layer whose output is fed to the hidden layer. The hidden layer is responsible for analysing the data, and the output of the hidden layer goes to the output layer. The number of hidden layers depends on the nature of the dataset and problem under consideration. If the neural network has multiple hidden layers, then such a neural network is termed as a deep neural network. The training algorithm for such a deep neural network is often termed as deep learning which is a subset of machine learning. Typically, the multiple hidden layers are responsible for computation of different levels of features of the data.

Long Short Term Memory (LSTM):

The LSTM networks are a specialized type of recurrent neural network (RNN) designed to process and predict data sequences by learning long-term dependencies. Unlike traditional RNNs, which suffer from vanishing or exploding gradient problems during training, LSTMs incorporate a unique architecture with gates and memory cells that help retain important information over long periods [9].

The LSTM primarily has 3 gates:
1) Input gate: This gate collects the presents inputs and also considers the past outputs as the inputs.
2) Output gate: This gate combines all cell states and produces the output.
3) Forget gate: This is an extremely important feature of the LSTM which received a cell state value governing the amount of data to be remembered and forgotten.

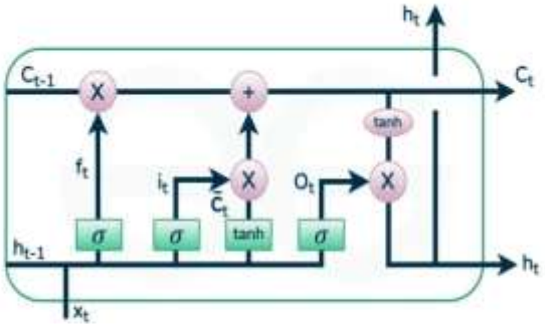

Fig.5 The LSTM Model

Figure 5 depicts the LSTM model.

The relation to forget by the forget gate is given by:

$$\boldsymbol{f = \sigma(W_f[h_{t-1}, x_t] + b_i)} \qquad \textbf{(6)}$$

Here,

f denotes forget gate activation

w_f are forget gate weights.

h_{t-1} Denotes Hidden state from the previous time step

x_t is present input.

b_i is the bias

The advantages of LSM are:

Capturing Long-Term Dependencies: LSTMs maintain long-term memory using the cell state, unlike traditional RNNs.

Mitigating Vanishing/Exploding Gradients: Gates help regulate gradient flow, enabling stable training over long sequences.

Versatility: Useful for several time series prediction problems.

However, the major challenge happens to be the problem of overfitting [10].

Convolutional Neural Networks (CNNs): The family of CNNs are the backbone of modern satellite object detection. CNNs automatically learn hierarchical features from raw images, eliminating the need for manual feature extraction. The Convolutional Neural Networks (CNNs) can automatically extract hierarchical characteristics from images, they have become the mainstay for image classification applications. These neural networks perform exceptionally well in applications like picture identification because they are specifically made for processing organised grid data.

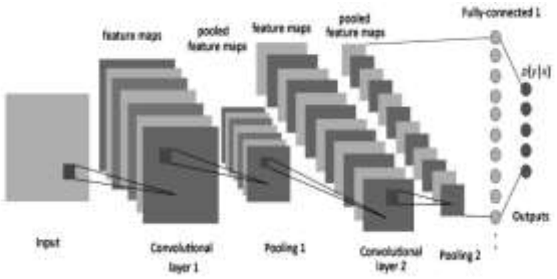

Fig.6 The CNN Model

Figure 6 depicts the CNN model.

The convolution operation is given by:

$$\boldsymbol{x(t) * h(t) = \int_{-\infty}^{\infty} x(\tau)h(t-\tau)d\tau} \qquad \textbf{(7)}$$

Here,

x(t) is the input

h(t) is the system under consideration.

y is the output

*is the convolution operation in continuous domain

For a discrete or digital counterpart of the data sequence, the convolution is computed using:

$$\boldsymbol{y(n) = \sum_{-\infty}^{\infty} x(k)h(n-k)} \qquad \textbf{(8)}$$

Here

x(n) is the input

h(n) is the system under consideration.

y is the output

*is the convolution operation in discrete domain.

Convolutional, pooling, and fully linked layers are among the layers that make up a CNN's architecture. Convolutional layers identify patterns in the input image by applying filters, hence identifying local features. By reducing spatial dimensions, pooling layers preserve significant information. High-level features are integrated for categorization in fully connected layers.

III. PREVIOUS WORK

This section presents the noteworthy contribution in the domain of research.

Thekkekara et al. [11] proposed a deep learning architecture with an attention mechanism on CNN-BiLSTM (CBA) and provide a comparative analysis to benchmark well-known deep learning models using the public dataset namely CLEF2017. Authors found that along with F1 score, precision and recall it is also vital to consider the Area under the curve - Receiver operating characteristic curve (AUC-ROC) and Mathews Correlation Coefficient (MCC) metrics for evaluating depression classification models since the MCC considers all the four values of a confusion matrix. Based on our experiments, the CBA model outperforms the existing state of the art model with an overall accuracy of 96.71% and scores of 0.85 and 0.77 for AUC-ROC and MCC, respectively.

Marriwala et al. [12] proposed the CNN and bi-LSTM models separately for depression detection. It is also observed that Bi-LSTM has better learning rate as compared to other models with accuracy 88% and validation accuracy 78%. There are some parameters such as precision, F1-score, recall and support are found for evaluation of models. In results, graphs for training loss, validation loss, training accuracy and validation accuracy are plotted. At last, by using confusion matrix depression can be detected for textual CNN Model, audio CNN model, LSTM model and Bi-LSTM against true label and predicted label.

Uddin et al. [13] employed recurrent neural networks, particularly LSTM models, to identify depressive symptoms from large-scale social media text. Their model achieved an accuracy of 84.2% and an F1-score of 0.83, demonstrating the ability of sequential deep learning models to capture long-term linguistic dependencies.

Trotzek et al. [14] analyzed Reddit posts using CNN and LSTM architectures with pretrained word embeddings. Their best-performing CNN model achieved an F1-score of 0.86 and an AUC of 0.88, outperforming traditional machine learning classifiers.

Chiong et al. [15] proposed an ensemble of machine learning classifiers including SVM and Random Forest for depression detection, reporting an accuracy of 79.6% and highlighting the effectiveness of linguistic and sentiment-based features.

Yang et al. [16] utilized BERT-based contextual embeddings for depression classification on Twitter data. Their transformer-based model achieved an accuracy of 87.9% and an AUC of 0.91, showing strong semantic modeling capabilities.

Khan et al. [17] compared classical ML models with deep neural networks on multilingual social media datasets. The BiLSTM model achieved an F1-score of 0.85, significantly outperforming Naïve Bayes and logistic regression.

Shen et al. [18] introduced an attention-based BiLSTM framework that improved interpretability while achieving an accuracy of 88.3% and F1-score of 0.87.

Kumar [19] proposed a hybrid deep learning framework combining CNN and LSTM architectures. The model achieved 89.1% accuracy and an F1-score of 0.88, demonstrating robust feature extraction.

Zhang et al. [20] employed RoBERTa for depression detection and reported an AUC of 0.93 and F1-score of 0.90, highlighting the superiority of transformer-based architectures.

Qasim et al. [21] focused on detecting depression severity levels using transformer models. Their approach achieved a macro F1-score of 0.82 across multiple severity classes.

Kim et al. [22] investigated LLM-derived embeddings for interpretable depression detection, achieving an accuracy of 86.5% and AUC of 0.89.

III. EVALUATION PARAMETERS

The classification accuracy is measured based on [23]:

1. **True Positive (TP):** It is indicative of the true or correct cases of the data to be in a particular class.
2. **True Negative (TN):** It is indicative of the true or correct cases of the data not to be in a particular class.
3. **False Positive (FP):** It is indicative of the false or incorrect cases of the data to be in a particular class.
4. **False Negative (FN):** It is indicative of the false or incorrect cases of the data not to be in a particular class.

Accuracy (Ac): It is an indicative of the accuracy of classification of the algorithm for data classification. Mathematically its defined as:

$$A_c = \frac{TP+TN}{TP+TN+FP+FN} \tag{9}$$

The future of depression detection lies in "multimodal fusion," where AI synthesizes data from multiple sources—such as wearable heart-rate monitors, voice recordings, and text logs—to create a holistic view of a patient's mental state. Rather than replacing human professionals, these models serve as powerful decision-support systems that enable "just-in-time" interventions [24]

Conclusion: The widespread adoption of social media platforms has resulted in large volumes of user-generated textual data, providing new opportunities for early identification of mental health conditions such as depression. Machine learning (ML) and deep learning (DL) techniques have emerged as effective tools for analyzing linguistic, semantic, and emotional patterns in social media posts. This paper presents a

comprehensive review of ML and DL approaches for depression detection from social media text. As these technologies move from the lab to real-world applications, they hold the promise of a more proactive healthcare system where mental health crises can be predicted and prevented before they fully manifest.

References:

1. P. K. Sonar and S. K. Yadav, "Early Detection of Depression Indication from Social Media Analysis," in ITM Web of Conferences, vol. 40, p. 03029, 2021.
2. A. Gupta, M. Sharma, and S. Kumar, "Depression Detection in Social Media: A Comprehensive Review of Machine Learning and Deep Learning Techniques," IEEE Access, vol. 13, pp. 12789-12815, Jan. 2025.
3. T. Siddiqui and A. Pandey, "A Comparative Study of Deep Learning Methods for Depression Detection in Social Media Data," Journal of Artificial Intelligence Research & Advances, vol. 12, no. 2, pp. 88-102, July 2025.
4. R. Chiong, G. S. Budhi, and S. Dhakal, "A textual-based featuring approach for depression detection using machine Learning classifiers and social media texts," Comput. Biol. Med., vol. 135, Aug. 2021.
5. B. Das, "An automated system of sentiment analysis from Bangla text using supervised learning techniques," in Proc. 4th IEEE Int. Conf. Comput. Commun. Syst. (ICCCS), 2019, pp. 43–47.
6. S. J. Pan et al., "Understanding emotions in text using deep Learning and big data," Comput. Hum. Behav., vol. 93, pp. 15–25, 2019
7. H. S. Alsagri and M. Ykhlef, "Machine Learning-based Approach for Depression Detection in Twitter Using Content and Activity Features," IEICE Trans. Inf. Syst., vol. E103-D, no. 8, pp. 1825–1832, 2020.
8. A. S. Sabitha, N. J. Gupta, and A. S. Shukla, "Predicting anxiety, depression and stress in modern life using machine Learning algorithms," Procedia Comput. Sci., vol. 173, pp. 333–342, 2020.
9. M. M. Tadesse, H. Lin, and B. Xu, "Detection of Depression-Related Posts from Social Media Using Machine Learning," IEEE Access, vol. 8, pp. 225346–225358, 2020.
10. K. S. Kumar, G. Geetha, and K. S. Sankaran, "Depression Detection in Social Media using Machine Learning Techniques," in Proc. Int. Conf. Adv. Comput. Commun. Syst. (ICACCS), 2021.
11. JP Thekkekara, S Yongchareon, V Liesaputr., " An attention-based CNN-BiLSTM model for depression detection on social media text," Expert Systems with Applications, Elsevier 2024, vol.249, 123834
12. N Marriwala, D Chaudhary, "A hybrid model for depression detection using deep learning," Measurement: Sensors, Elsevier 2023, vol.25, 100587.
13. M. Z. Uddin, M. M. Hassan, A. Alsanad, and A. Alqurashi, “A deep learning-based early warning system for depression detection using social media text,” Neural Computing and Applications, vol. 34, no. 2, pp. 721–744, 2021.
14. M. Trotzek, S. Koitka, and C. M. Friedrich, “Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences,” IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 3, pp. 588–601, 2021.
15. Y. Chiong, L. Budhi, and D. B. Pham, “Depression detection using sentiment and linguistic features from social media,” Expert Systems with Applications, vol. 186, p. 115759, 2022.
16. Y. Yang, J. Liu, K. Zhang, and Y. Wang, “Depression detection on social media using BERT-based deep learning,” IEEE Access, vol. 10, pp. 18432–18443, 2022.
17. S. Khan, A. Hussain, and M. Afzal, “Comparative analysis of machine learning and deep learning techniques for depression detection on multilingual social media data,” Computers in Biology and Medicine, vol. 151, p. 106236, 2023.
18. L. Shen, J. Chen, and X. Liu, “Attention-based BiLSTM for depression detection from social media text,” Information Processing & Management, vol. 60, no. 1, p. 102912, 2023.
19. T. S. Kumar, “A hybrid deep learning framework for automatic depression detection from social media posts,” International Journal of Intelligent Systems and Applications in Engineering, vol. 12, no. 4, pp. 3217–3231, 2024.
20. Z. Zhang, H. Wang, and Y. Liu, “Depression detection from social media text using RoBERTa,” IEEE Access, vol. 12, pp. 44115–44127, 2024.

21. A. Qasim, G. Mehak, N. Hussain, A. Gelbukh, and G. Sidorov, "Detection of depression severity in social media text using transformer-based models," Information, vol. 16, no. 2, p. 114, 2025.

22. S. Kim, O. Imieye, and Y. Yin, "Interpretable depression detection from social media text using large language model embeddings," arXiv preprint arXiv:2506.06616, 2025.

23. M. Nadeem, S. Rauf, and A. Iqbal, "Early depression detection from social media using BERT–BiLSTM hybrid models," IEEE Access, vol. 13, pp. 55210–55223, 2025

24. L. Wang and J. Wu, "Advances in artificial intelligence-based depression diagnosis: A systematic review," ICCK Transactions on Emerging Topics in Artificial Intelligence, vol. 2, no. 3, pp. 148–156, 2025.