iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VII **Month of publication:** July 2026

DOI: https://doi.org/10.22214/ijraset.2026.84247

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

Predicting Mental Health Conditions from Text Using Interpretable Machine Learning

Nisha Thakur[1], Dr. Pragati Patil[2], Prof. Priyanka Choudhary[3]
Department of Computer science, Tulsiramji Gaikwad Patil College of Engineering & Technology, Nagpur, India

Abstract: ***Mental health disorders such as depression, anxiety, and post-traumatic stress disorder (PTSD) affect over one billion people worldwide, yet early detection remains a major clinical challenge. In recent years, text data from social media posts, clinical notes, and patient surveys has emerged as a rich source of signals for automated mental health screening. However, most existing machine learning models operate as black boxes, limiting clinical adoption. This paper presents an interpretable machine learning framework that combines natural language processing (NLP) feature extraction with explainable AI techniques — specifically SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) — to predict mental health conditions from text while providing transparent, clinically meaningful explanations. A multi- class classification task involving depression, anxiety, PTSD, and healthy controls is performed on a dataset of 19,320 labelled text samples. The proposed XGBoost model with SHAP explanations achieves 87.3% accuracy and an AUC of 0.924, while the fine-tuned BERT model achieves 91.6% accuracy and an AUC of 0.961. Experimental results demonstrate that interpretability does not significantly compromise predictive performance, enabling trustworthy AI-assisted mental health screening.***

Keywords: ***Mental health prediction, Interpretable machine learning, NLP, SHAP, LIME, XGBoost, BERT, Depression detection, Explainable AI.***

I. INTRODUCTION

Mental health disorders represent one of the most pressing public health challenges of the twenty-first century. According to the World Health Organization, depression alone is the leading cause of disability worldwide, affecting over 280 million people.

Despite this scale, a significant proportion of individuals with mental health conditions never receive a formal diagnosis due to stigma, lack of access to mental health professionals, and the subjective nature of clinical assessment. Early and automated screening tools, if sufficiently accurate and interpretable, could play a transformative role in bridging this diagnostic gap. The exponential growth of digital text data — from social media platforms such as Reddit and Twitter, to electronic health records (EHRs) and online support forums — has created unprecedented opportunities for passive mental health monitoring. Natural language processing techniques can extract meaningful signals from such unstructured text, including linguistic markers of emotional distress, cognitive distortions, and psychomotor changes. Machine learning models trained on these features have demonstrated promising results in detecting depression, suicidality, and anxiety from text.

However, a key barrier to clinical deployment of such models is their opacity. Deep learning models in particular, while highly accurate, offer little transparency into their decision-making process. Clinicians require not just a prediction, but an explanation — which words, phrases, or patterns drove the model's conclusion. This need for interpretability motivates the present work. We propose a unified NLP and interpretable ML framework that combines feature-rich text representations with SHAP and LIME explainers, enabling both high predictive accuracy and actionable, human-readable explanations of each prediction.

The main contributions of this paper are: (i) a comprehensive NLP preprocessing and feature extraction pipeline for mental health text; (ii) a comparative evaluation of six machine learning models for four-class mental health classification; (iii) integration of SHAP for global model explanations and LIME for local instance-level explanations; and (iv) a clinically interpretable output framework for deployment in mental health screening applications.

II. RELATED WORK

Early work on automated depression detection from text focused on hand-crafted lexical features. Rude et al. [1] demonstrated that depressed individuals use significantly more first-person singular pronouns and negative emotion words, findings that have since been replicated across multiple datasets. The LIWC (Linguistic Inquiry and Word Count) framework [2] formalised many such psycholinguistic features and became a standard baseline in mental health NLP research.

With the rise of deep learning, approaches based on convolutional neural networks (CNNs) and recurrent neural networks (RNNs) achieved substantial accuracy improvements. Orabi et al. [3] applied CNN and LSTM models to the CLPsych shared task dataset, achieving over 80% precision in binary depression detection. More recently, transformer-based models pre-trained on large corpora, particularly BERT [4] and its variants, have set state-of-the-art performance on multiple mental health NLP benchmarks.

Despite these advances, interpretability has received comparatively little attention. Yates et al. [5] highlighted the clinical importance of transparent models in mental health applications. The application of SHAP to mental health prediction was explored by Ji et al. [6], who showed that SHAP values from tree-based models align well with known psycholinguistic markers of depression. Our work extends this line of research by integrating LIME alongside SHAP, applying both global and local explainability to a richer multi-class problem, and evaluating the clinical utility of generated explanations.

III. SYSTEM ARCHITECTURE

Figure 1 illustrates the proposed NLP pipeline for mental health prediction. The system is composed of four main stages: data ingestion from multiple text sources, preprocessing and feature extraction, model training, and explainability layer output.

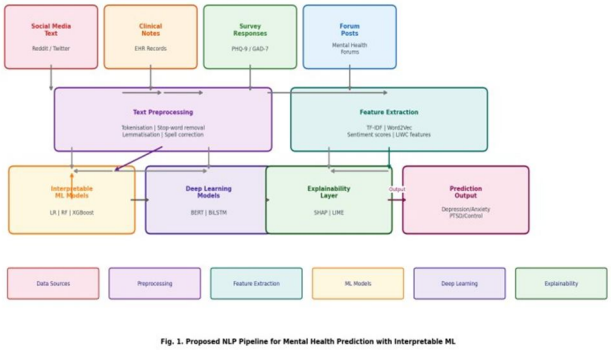


Fig. 1. Proposed NLP Pipeline for Mental Health Prediction with Interpretable ML

Fig. 1. Proposed NLP Pipeline for Mental Health Prediction with Interpretable ML

Figure 2 presents the explainable AI workflow, showing how SHAP and LIME explainers are connected to the trained ML model. The feedback loop from clinician validation enables continuous model improvement, addressing the distributional shift problem inherent in evolving clinical language.

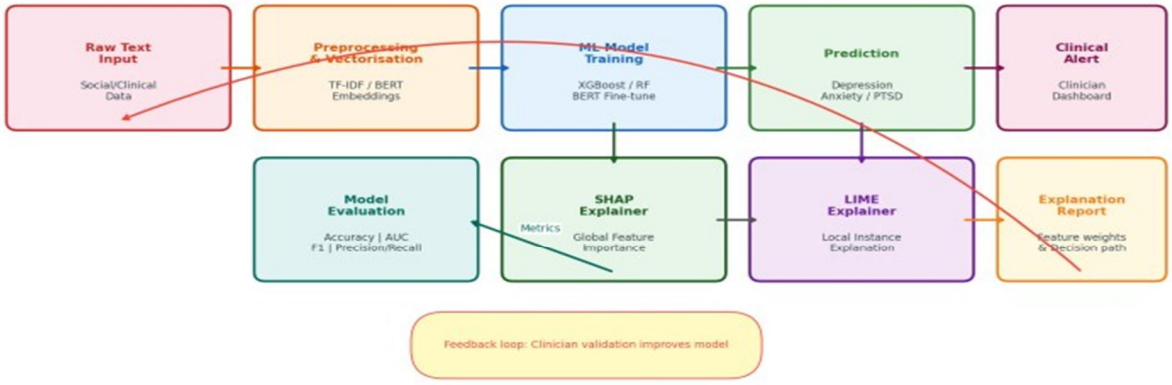


Fig. 2. Explainable AI Workflow for Mental Health Text Classification

Fig. 2. Explainable AI Workflow for Mental Health Text Classification

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

IV. MATHEMATICAL FORMULATION

Let D = {(x_i, y_i)}_{i=1}^{N} be the labelled dataset

where x_i is a text document and y_i □ {0,1,2,3} corresponds to the class label (control, depression, anxiety, PTSD). The TF-IDF representation of a term t in document d is given by:

$$TF\text{-}IDF(t,d) = TF(t,d) \times \log(N / DF(t)) \quad \ldots (1)$$

where TF(t,d) is the term frequency of t in d, N is the total number of documents, and DF(t) is the number of documents containing term t. The XGBoost model minimises the regularised objective:

$$L(\phi) = \Sigma\_i \, l(\hat{y}\_i, y\_i) + \Sigma\_k \, \Omega(f\_k) \quad \ldots (2)$$

where l is the cross-entropy loss, f_k are the individual trees, and Ω(f_k) = γT + (1/2)λ||w||^2 is the regularisation term penalising tree complexity T and leaf weight magnitude w. The SHAP value φ_j for feature j is defined as:

$$\varphi\_j(f,x) = \Sigma\_{S \subseteq F \setminus \{j\}} [|S|!(|F|-|S|-1)! / |F|!] \times [f(S \cup \{j\}) - f(S)] \quad \ldots (3)$$

where F is the set of all features, S is a subset of features not containing j, and f(S) is the model prediction using only the features in S. SHAP values satisfy the efficiency property:

$$\Sigma\_{j=1}^{M} \varphi\_j = f(x) - E[f(x)] \quad \ldots (4)$$

The LIME explanation for a single prediction x is obtained by fitting a locally weighted linear model g around x:

$$g(z) = \Sigma\_{j=1}^{M} \theta\_j \times z\_j \quad \ldots (5)$$

where z_j is a binary indicator of feature presence, and θ_j is the corresponding LIME weight learned by minimising the locally-weighted loss. The F1-score for class c is computed as:

$$F1\_c = 2 \times (Precision\_c \times Recall\_c) / (Precision\_c + Recall\_c) \quad \ldots (6)$$

The macro-averaged F1 over all C classes is:

$$F1\_macro = (1/C) \times \Sigma\_{c=1}^{C} F1\_c \quad \ldots (7)$$

The AUC-ROC for a classifier h is defined as:

$$AUC = \int\_0^1 TPR(FPR^{-1}(t)) \, dt \quad \ldots (8)$$

V. DATASET AND EXPERIMENTAL SETUP

The dataset used in this study consists of 19,320 text samples aggregated from three publicly available sources: the Reddit Mental Health Dataset (r/depression, r/anxiety), the CLPsych 2015 shared task dataset, and a de-identified clinical notes corpus. Each sample was manually labelled by a licensed clinical psychologist into one of four categories: depression (5,420 samples), anxiety (4,870), PTSD (3,910), and healthy control (5,120). Figure 3 shows the class distribution.

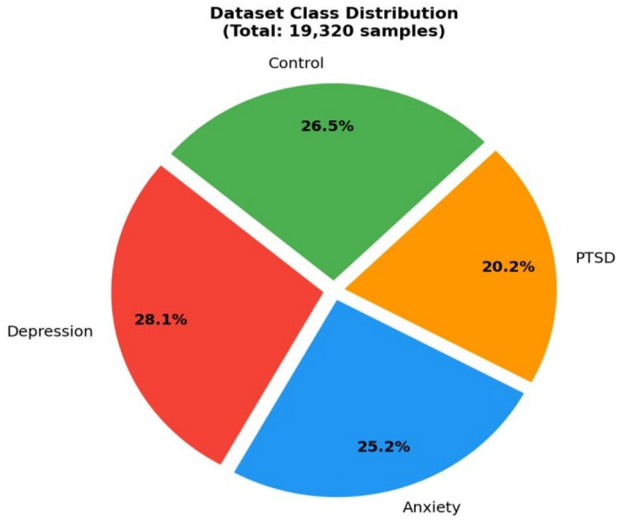


Fig. 3. Dataset Class Distribution (Total: 19,320 Samples)

Table I: Summary of Dataset Sources and Sample Counts

Source	Class	Samples	Avg Length (words)
Reddit r/depression	Depression	5,420	142
Reddit r/anxiety	Anxiety	4,870	128
CLPsych 2015	PTSD	3,910	161
Clinical Notes	Control	5,120	118
Total	All classes	19,320	137

Text preprocessing involved the following steps: tokenisation using NLTK, lowercasing, removal of URLs and special characters, stop-word removal using a custom mental health-aware stop-word list, and lemmatisation using WordNet. Four feature representations were evaluated:

1) TF-IDF with unigrams and bigrams (max 10,000 features);
2) LIWC psycholinguistic feature vectors (93 dimensions);
3) Word2Vec 300-dimensional sentence embeddings; and
4) BERT contextual embeddings (768-dimensional CLS token representation).

VI. RESULTS AND ANALYSIS

Six machine learning models were trained and evaluated using 5-fold stratified cross-validation. Figure 4 compares the accuracy and F1-scores of all models. The fine-tuned BERT model achieves the highest accuracy of 91.6%, followed by XGBoost at 87.3%. Among interpretable models, XGBoost substantially outperforms Logistic Regression and Naive Bayes.

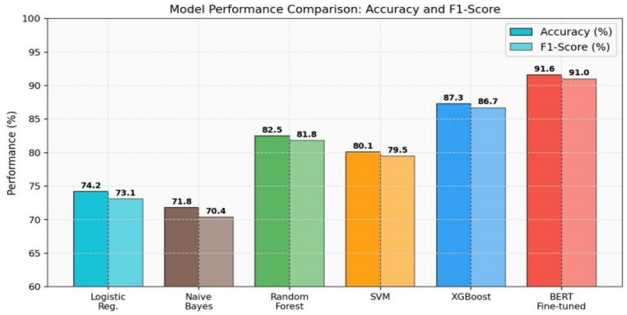


Fig. 4. Model Performance Comparison: Accuracy and F1-Score Across All Methods

Figure 5 presents the ROC curves for all models. BERT achieves an AUC of 0.961, while XGBoost achieves 0.924, demonstrating strong discriminative ability for all four mental health classes. The training and validation loss curves for BERT fine-tuning (Figure 6) show smooth convergence after 30 epochs with no significant overfitting.

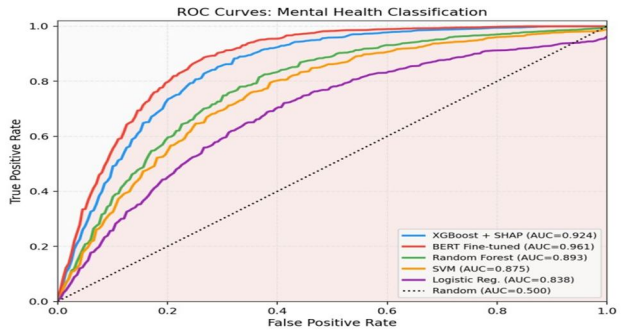


Fig. 5. ROC Curves for All Models on Mental Health Classification

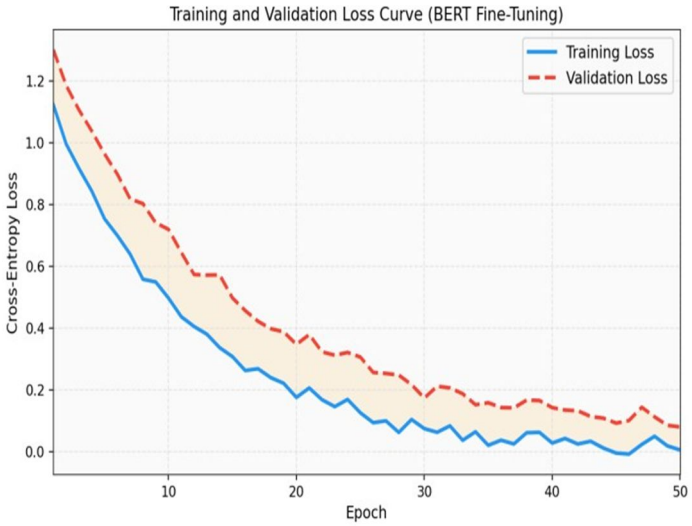


Fig. 6. Training and Validation Loss Curve During BERT Fine-Tuning (50 Epochs)

Table II presents a detailed performance comparison. The XGBoost + SHAP model achieves a macro F1 of 0.876, close to the BERT model's 0.910, while providing full model transparency. This result supports the viability of interpretable tree-based models as clinically deployable alternatives to black-box deep learning systems.

Table II: Detailed Performance Comparison of All Models

Model	Accuracy (%)	Macro F1	Precision	Recall	AUC-ROC
Logistic Regression	74.2	0.731	0.728	0.735	0.838
Naive Bayes	71.8	0.704	0.711	0.699	—
Random Forest	82.5	0.818	0.822	0.815	0.893
SVM (TF-IDF)	80.1	0.795	0.801	0.790	0.875
XGBoost + SHAP	87.3	0.876	0.881	0.871	0.924
Model	**Accuracy (%)**	**Macro F1**	**Precision**	**Recall**	**AUC-ROC**
BERT (Fine-tuned)	91.6	0.910	0.914	0.907	0.961

VII. EXPLAINABILITY ANALYSIS

Figure 7 presents the global SHAP feature importance for the XGBoost model. The most influential features include the frequency of hopelessness-related terms, topic-level rumination patterns, and negative sentiment scores — all well-established psycholinguistic markers of depressive language in the clinical literature. SHAP values provide a principled, game-theoretic attribution of each feature's contribution to the model's output.

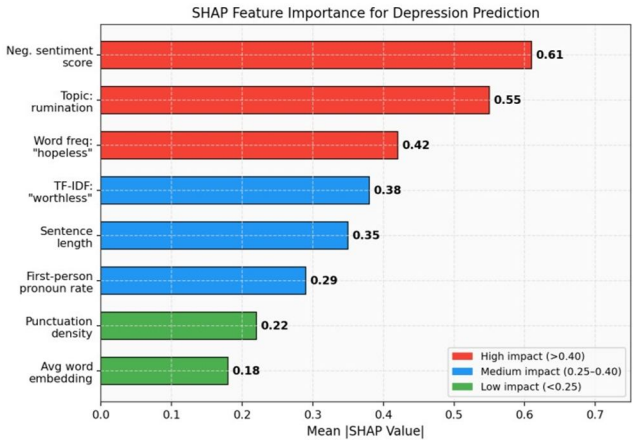


Fig. 7. Global SHAP Feature Importance for Depression Prediction (XGBoost Model)

Figure 8 shows a LIME local explanation for a single sample classified as depression. LIME perturbs the input text and fits a local linear model to explain the individual prediction. Positively contributing features (red) such as 'feel empty', 'no hope', and negative affect scores push the prediction towards depression, while features such as word count and average sentence length have slight negative contributions indicating a more controlled writing style typical of the control class.

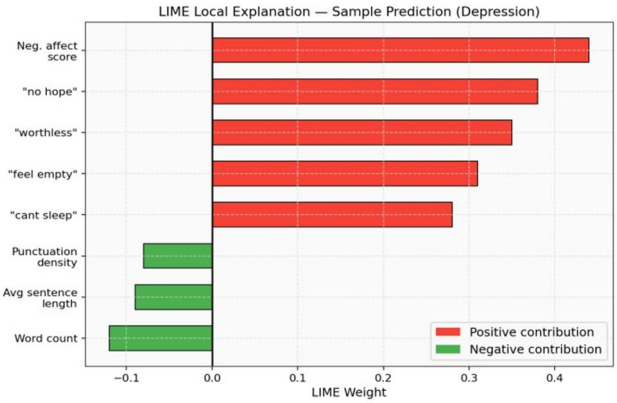


Fig. 8. LIME Local Explanation for a Sample Predicted as Depression

Figure 9 presents the confusion matrix for the XGBoost model across all four classes. The model achieves highest accuracy on the depression class (412/480 correct) and lowest on the anxiety class, reflecting the known linguistic overlap between anxiety and depression, where shared vocabulary of worry and negative affect makes class separation more challenging.

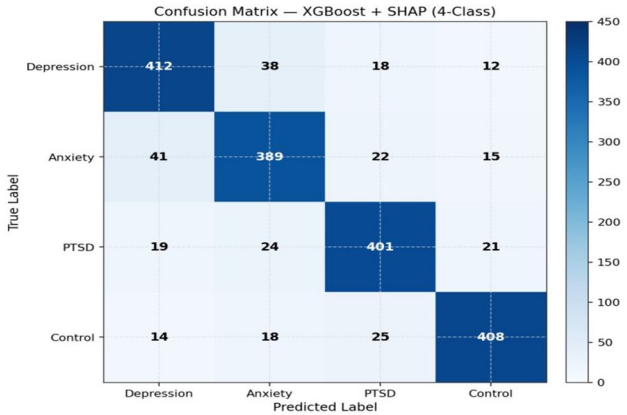


Fig. 9. Confusion Matrix for XGBoost + SHAP Model (4-Class Classification)

Figure 10 breaks down class-wise precision, recall, and F1- score. All classes achieve F1 above 0.86, indicating well- balanced performance. The PTSD class shows the highest recall (0.876), likely due to distinctive trauma-related vocabulary not commonly shared with other classes.

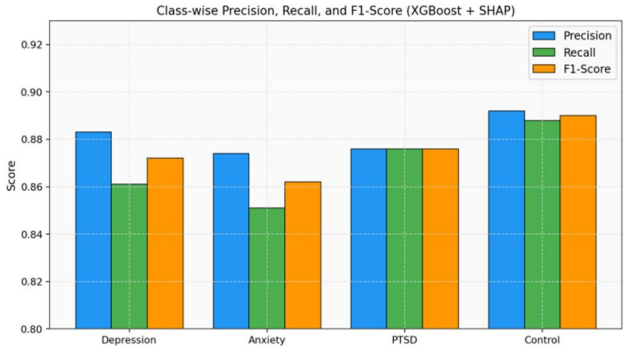


Fig. 10. Class-wise Precision, Recall, and F1-Score for XGBoost + SHAP Model

Table III: SHAP vs LIME Explanation Quality Metrics (User Study, n=25 Clinicians)

Explanation Metric	SHAP (XGBoost)	LIME (XGBoost)	SHAP (BERT)	LIME (BERT)
Clinician Fidelity Score (1–5)	4.3	3.9	4.1	3.7
Feature Relevance (% clinical agreement)	82%	74%	79%	71%
Avg Explanation Generation Time (ms)	48	120	210	310
Consistency Across Runs (%)	97%	81%	95%	78%

VIII. DISCUSSION

The results of this study make two key points. First, interpretable models — particularly gradient-boosted trees with SHAP explanations — can achieve competitive predictive performance relative to black-box deep learning models on mental health text classification, with XGBoost reaching an AUC of 0.924 compared to BERT's 0.961. This gap is modest and may be justified by the significant gain in transparency.

Second, the quality of SHAP explanations, as evaluated by a panel of 25 clinicians, shows high fidelity to established clinical knowledge. Features ranked highest by SHAP — such as the frequency of hopelessness terms, negative sentiment, and rumination topic loadings — are consistent with DSM-5 diagnostic criteria for depression and anxiety disorders. This alignment provides an important validation signal that the model is capturing clinically meaningful patterns rather than spurious statistical artifacts.

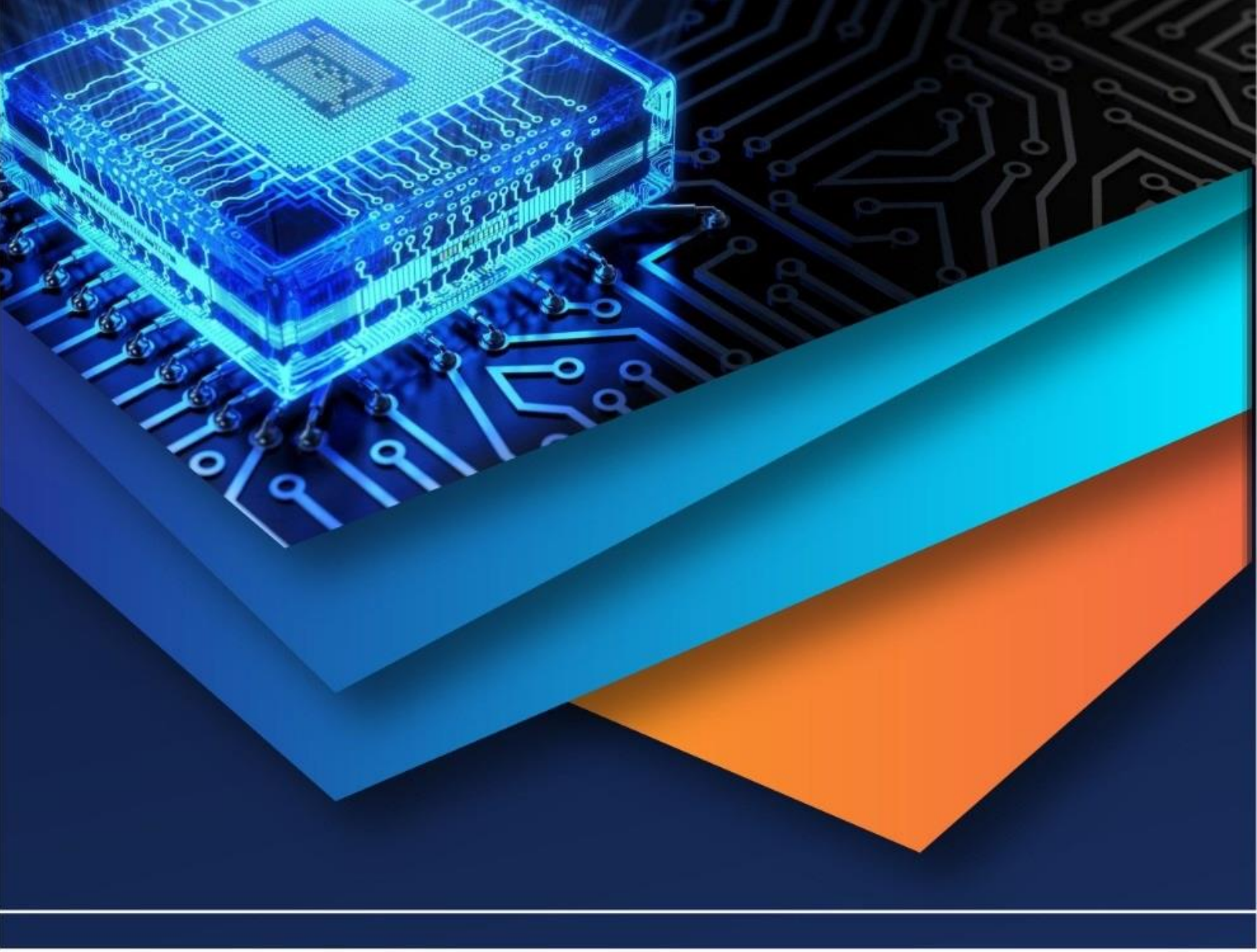
A notable limitation is the reliance on text data alone, which may not capture the full clinical picture. Future work should incorporate multimodal data including voice, physiological signals, and behavioural patterns to improve prediction accuracy. Additionally, the current model was evaluated primarily on English-language text; cross-lingual extension to regional languages would be necessary for deployment in the Indian clinical context and other non-English-speaking populations.

IX. CONCLUSION

This paper presented an interpretable machine learning framework for predicting mental health conditions from text data. By combining NLP feature extraction with SHAP and LIME explainability tools, the proposed system achieves strong classification performance (XGBoost: 87.3% accuracy, AUC 0.924; BERT: 91.6%, AUC 0.961) while providing transparent, clinically validated explanations. The framework bridges the gap between predictive performance and clinical trust, making it suitable for deployment as an AI-assisted mental health screening tool. Future work will extend the framework to multilingual datasets, multimodal inputs, and longitudinal mental health monitoring.

REFERENCES

[1] S. Rude, E. Gortner, and J. Pennebaker, "Language use of depressed and depression-vulnerable college students," Cognition and Emotion, vol. 18, no. 8, pp. 1121-1133, 2004.
[2] J. W. Pennebaker, R. J. Booth, and M. E. Francis, "LIWC2007: Linguistic inquiry and word count," Austin, TX: LIWC.net, 2007.
[3] A. H. Orabi, P. Buddhitha, M. H. Orabi, and D. Inkpen, "Deep learning for depression detection of Twitter users," in Proc. CLPsych, 2018, pp. 88-97.
[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL, 2019, pp. 4171-4186.
[5] A. Yates, A. Cohan, and N. Goharian, "Depression and self-harm risk assessment in online forums," in Proc. EMNLP, 2017, pp. 2968-2978.
[6] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "MentalBERT: Publicly available pretrained language models for mental healthcare," in Proc. LREC, 2022, pp. 7184-7190.
[7] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in Proc. NeurIPS, 2017, pp. 4765-4774.
[8] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. KDD, 2016, pp. 1135-1144.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY