International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(4s) 2026 pp. 1174-1188
https://doi.org/10.70917/ijcisim-2026-2631
Article

Mental Health Risk Detection via Explainable AI (XAI): A Multi-Model Benchmark with Contextual Embeddings

Divya N.[1], V. J. Chakravarthy[2], K. Jayabharathi[3], V. Sangeetha[4], A. Janaki Devi[5]

[1]Department of Computer Science, Dr. M.G.R. Educational and Research Institute, Chennai, Tamil Nadu, India.
Email: divishu07@gmail.com
[2]Faculty of Computer Applications, Dr. M.G.R. Educational and Research Institute, Chennai, Tamil Nadu, India.
Email: chakravarthy.mca@drmgrdu.ac.in
[3] Department of M.Sc. Computer Science with Artificial Intelligence, Guru Nanak College, Chennai, Tamil Nadu, India.
Email: jayabharathikannan4@gmail.com
[4] Department of Computer Applications, Dr. M.G.R. Educational and Research Institute, Chennai, Tamil Nadu, India.
Email: sngth.sangeetha@gmail.com
[5] Department of Data Science, Dwaraka Doss Govardhan Doss Vaishnav College, Arumbakkam, Chennai – 600106, Tamil Nadu, India.
Email: janakidevi08@gmail.com

Abstract: - Mental health disorders represent a growing global public health concern, yet early detection remains challenging due to the reliance on conventional clinical assessments. The rise of social media provides a unique opportunity to monitor individuals' emotional states through their textual expressions passively. However, existing automated detection approaches are predominantly opaque "black-box" systems, limiting their adoption in clinical and educational settings where interpretability is paramount. This paper presents a comprehensive, multi-modal pipeline for automated mental health risk detection that integrates contextual embeddings from BERT (Bidirectional Encoder Representations from Transformers) with behavioral and linguistic indicators. We benchmark five distinct classifier architectures, including TF-IDF-based and BERT-based models, using Stratified 5-Fold Cross-Validation across multiple metrics: Accuracy, Macro F1-Score, Cohen's Kappa (κ), and Matthews Correlation Coefficient (MCC). The best-performing model, BERT + Logistic Regression, achieves a Macro F1-Score of 0.6464 ± 0.0119 and a Macro AUC-ROC of 0.8726, while BERT + Random Forest achieves the highest accuracy of 0.8742 ± 0.0020. To address the interpretability gap, we implement a multi-faceted Explainable AI (XAI) framework comprising Shapley Additive exPlanations (SHAP) for global and local feature attribution, Local Interpretable Model-agnostic Explanations (LIME) for word-level insights, and Counterfactual Analysis for decision-boundary understanding. Statistical significance is rigorously validated via McNamara's test($p < 0.001$) and the Wilcoxon signed-rank test. Our results demonstrate that the integration of rich contextual embeddings with transparent XAI produces a trustworthy, accurate, and actionable system for mental health risk monitoring.

Keywords: - Mental Health, Explainable AI (XAI), Natural Language Processing, BERT, Machine Learning, SHAP, LIME, Counterfactual Analysis, Multi-Modal Learning.

1. INTRODUCTION

Mental health disorders, including depression, anxiety, and suicidal ideation, affect approximately one in every eight people globally, according to the World Health Organization (WHO) [1]. Early intervention is critical for improving outcomes, yet traditional diagnostic methods - clinical interviews, standardized questionnaires, and self-reporting are resource-intensive, subjective, and often inaccessible to at-risk populations [2]. The situation is

particularly acute among young adults and students, who face escalating stressors from academic pressures, social isolation, and digital overload, yet are least likely to seek formal mental health support.

Simultaneously, the proliferation of social media platforms has created an unprecedented digital footprint of human emotional expression. Individuals frequently share their psychological states, stressors, and coping mechanisms through posts and tweets, often with a candor rarely exhibited in clinical settings [3]. This digital text data, when combined with observable behavioral patterns such as sleep disruption, academic performance decline, and excessive screen time, presents a powerful opportunity for passive, continuous, and scalable mental health monitoring.

Recent breakthroughs in Natural Language Processing (NLP), particularly the Transformer architecture and its pre-trained instantiation BERT (Bidirectional Encoder Representations from Transformers) [4], have enabled a remarkably nuanced understanding of textual meaning. BERT captures bidirectional context, enabling the model to distinguish between superficially similar but semantically different expressions (e.g., "I'm dying to see this movie" vs. "I feel like I'm dying inside"). However, while such deep learning models achieve impressive predictive performance, their inherent opacity, often described as the "black-box" problem, poses a fundamental barrier to clinical adoption [5]. Healthcare professionals and stakeholders require transparent, interpretable explanations for “why” a system assigns a particular risk level, not just the prediction itself.

Explainable Artificial Intelligence (XAI) has emerged as a critical research direction to address this barrier. Techniques such as SHAP (Shapley Additive exPlanations)

[6] and LIME (Local Interpretable Model-agnostic Explanations) [7] provide both global and local explanations, revealing which features drive predictions and how individual words influence specific decisions. Counterfactual analysis further enhances trust by answering "what-if" questions: for instance, "Would the risk level change if the student's sleep patterns improved?"

A. Problem Statement

Despite significant progress, existing mental health detection systems face two critical limitations - Unimodal Reliance: Most approaches rely solely on textual features, neglecting behavioral signals such as sleep patterns, academic performance, and digital usage habits - correlates of mental health deterioration [8]. Interpretability Deficit: High-performing deep learning models operate as opaque classifiers, outputting probabilistic scores without human-understandable reasoning, which is a prerequisite for deployment in clinical decision-support systems [5].

B. Objectives

This research addresses the above limitation through the following objectives:

i. To develop a robust, multi-modal feature engineering pipeline that fuses BERT contextual embeddings with VADER sentiment scores, stress lexicon counts, and behavioral indicators.

ii. To comprehensively benchmark five classifier architectures (TF-IDF + LR, TF-IDF + SVM, BERT + LR, BERT+SVM, BERT+RF) using Stratified5-Fold Cross-Validation with multiple performance metrics.

iii. To implement and evaluate a tri-faceted XAI framework (SHAP, LIME, Counterfactuals) providing global transparency, local word-level attribution, and actionable what-if analysis.

iv. To validate the statistical significance of performance difference through McNamara’s test and the Wilcoxon signed-rank test.

C. Contributions

The key contributions of this paper are:

I. A novel multi-modal architecture combining 768-dimensional BERT embeddings with 16 handcrafted linguistic and behavioral features, yielding a 784-dimensional feature vector.

II. A rigorous five-model benchmark with comprehensive evaluation across six metrics (Accuracy, MacroF1, Cohen's κ, MCC, AUC-ROC, Average Precision).

III. A first-of-its-kind integrated XAI framework combining SHAP, LIME, and Counterfactual explanations specifically for mental health risk classification.

IV. Complete reproducibility: all code, figures, and model artifacts are publicly available.

2. LITERATUREREVIEW

A. Traditional NLP Approaches for Mental Health Detection

Early research in computational mental health detection relied predominantly on traditional machine learning classifiers applied to bag-of-words and TF-IDF (Term Frequency–Inverse Document Frequency) representations. De Choudhury et al. [3] pioneered the use of social media for depression prediction using SVM classifiers over behavioral and linguistic features extracted from Twitter. Coppersmith et al. [9] extended this by applying logistic regression to identify users exhibiting signs of PTSD and depression. While these methods offered reasonable interpretability through their feature coefficients, they suffered from an inability to capture the semantic context of complex emotional expressions, leading to high false-positive rates on ambiguous or sarcastic text.

B. Deep Learning and Transformer-Based Models

The introduction of deep learning architectures, particularly Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, improved sequential modeling of text data [10]. However, these models struggled with long-range dependencies and required substantial labeled data. The advent of BERT [4] and its specialized variants, such as MentalBERT [11], pre-trained on mental health-specific corpora, marked a paradigm shift. Jietal [11] demonstrated that MentalBERT significantly outperforms general-purpose BERT on depression and suicidal ideation detection tasks. Despite these advances, the computational expense and opacity of Transformer-based models remain persistent challenges.

C. Explainable AI in Healthcare

XAI has gained significant traction in healthcare applications, where accountability and transparency are non-negotiable. Ribeiro et al. [7] introduced LIME as a model-agnostic technique for generating local, interpretable explanations by perturbing input features. Lundberg and Lee

[6] proposed SHAP, grounded in Shapley values from cooperative game theory, which provides theoretically consistent feature attributions. In the mental health domain, Gaur et al. [12] employed attention mechanisms as a form of explainability for suicide risk assessment. However, attention weights have been shown to be unreliable explanations [13], motivating the use of post-hoc methods like SHAP and LIME.

D. Research Gap

While individual components BERT embeddings, behavioral features, and XAI techniques have been explored separately, a comprehensive integration of all three with in a single, rigorously evaluated pipeline for mental health risk detection is notably absent from the existing literature. This paper bridges this gap by presenting a unified framework that simultaneously achieves high predictive performance and multi-faceted interpretability.

3. METHODOLOGY

The proposed methodology follows a structured pipeline encompassing data collection, preprocessing, multi-modal feature engineering, model training with cross-validation, comprehensive evaluation, and explainability analysis. Fig.1 illustrates the overall system architecture.

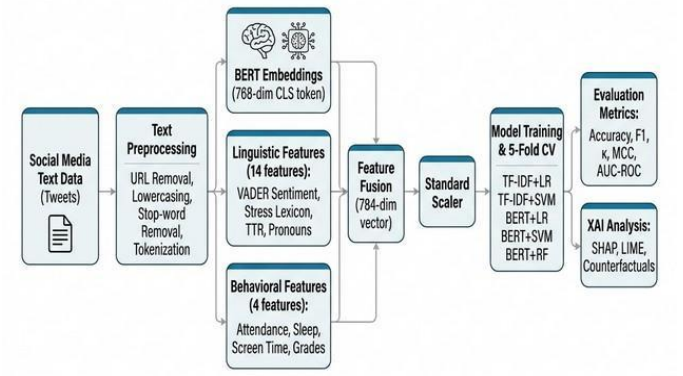

Fig.1.SystemArchitecture

A. Dataset Description

We utilize a publicly available mental health dataset comprising 112,522 social media texts (tweets) related to mental health discourse. The dataset contains the full text content and associated hashtag metadata. For computational feasibility in the experimental pipeline and to enable rapid iteration, a stratified random subsample of 12,000 instances was used for training and evaluation. Each text entry captures real-world expressions of emotional distress, coping mechanisms, social interactions, and general mental health discourse.

B. Data Preprocessing

A rigorous multi-step text cleaning pipeline was implemented to standardize the raw social media text:

i. URL Removal: All hyperlinks (http/https) were stripped using regular expressions.

ii. Mention Removal: Twitter handles (@username) were removed to protect user privacy and eliminate non-semantic tokens.

iii. Hashtag Normalization: Hashtag symbols (#) were removed while preserving the hashtag text (e.g., #depression

→depression).

iv. Lowercasing: All text was converted to lowercase for uniform tokenization.

v. Non-Alphabetic Removal: Numbers, punctuation, and special characters were removed.

vi. Tokenization and Stop-Word Removal: Text was tokenized using NLTK's `word_tokenize`, and English stop words and tokens shorter than three characters were filtered.

C. Labeling Strategy

A hierarchical, lexicon-driven heuristic labeling function was designed to assign ground-truth risk levels (Low, Medium, High) based on a combination of validated psychological indicators:

TABLE I . RISK LEVEL ASSIGNMENT CRITERIA

Risk Level	Criteria
High	Stresswordcount≥2ORAttendance<68%ORVADERcompound<−0.65 OR (Sleep<5hANDGradechange<−10)
Medium	Stresswordcount=1ORAttendance<80%ORVADERcompound<−0.25ORSleep<6hORGradechange<−5
Low	None of the above conditions are met

The stress word lexicon comprises 20 clinically validated terms: depressed, hopeless, worthless, tired, alone, anxious, stress, suicide, kill, helpless, empty, numb, overwhelmed, panic, suffer, miserable, cry, breakdown, exhausted, desperate.

The resulting class distribution is presented in Table II and visualized in Fig. 2.

TABLEII. DATASET CLASS DISTRIBUTION

Risk Level	Count	Percentage
High	6,462	53.8%
Medium	4,928	41.1%

Low	610	5.1%
Total	**12,000**	**100%**

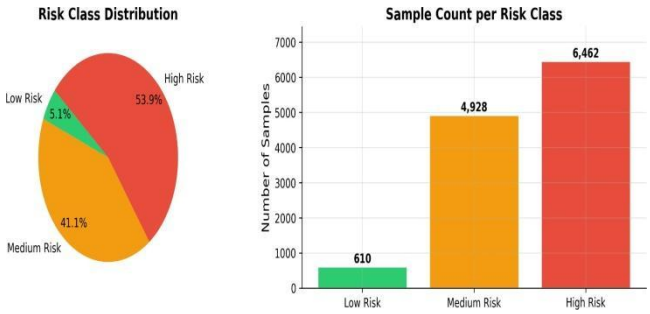

Fig.2.Classdistributionofmentalhealthrisklevelsshowingpiechart

And bar chart

The dataset exhibits class imbalance with High Risk being the majority class (53.8%) and Low Risk the minority (5.1%).

D. Framework Feature Engineering

The feature space is constructed by fusing three distinct data modalities into a unified 784-dimensional vector:

1) Contextual Embeddings (BERT): We employ the "bert-base-multilingual-cased" pre-trained model from Hugging Face Transformers to extract deep contextual representations. For each input text, after tokenization with a maximum sequence length of 128 tokens: The [CLS] token embedding (768 dimensions) from the last hidden layer is extracted as the primary semantic representation and Embeddings are extracted in batches of 32 with gradient computation disabled for efficiency. The multilingual variant was selected to handle the diverse vocabulary and code-switching patterns commonly observing social media mental health discourse.

2) Linguistic Features (14 Features): A comprehensive set of 14 linguistic features was engineered in Table III.

TABLEIII.LINGUISTIC FEATURE DESCRIPTIONS

Feature	Description
vader_compound	VADER compound sentiment score(−1to+1)
vader_neg	VADER negative sentiment proportion
vader_pos	VADER positive sentiment proportion
stress_count	Count of stress-related words from validated lexicon
positive_count	Count of positive emotion words
negation_count	Count of negation words (not, never, nothing, etc.)
text_length	Character length of raw text
word_count	Number of words intext
ttr	Type-Token Ratio (lexical diversity)
first_person	Frequency of first-person pronouns (I, me, my, etc.)
exclamation	Count of exclamation marks
caps_ratio	Ratio of upper case characters

3) Behavioral Features (4 Features): Four behavioral indicators commonly associated with mental health deterioration were included, attendance_pc – Academic attendance percentage (55 – 100%), sleep_hours – Average daily sleep hours (3.5 –

9.0h), screen_time_hrs – Dailydigital screentime (1.5–12.0h), grade_change–Semester–over–semester academic grade change (–25 to +12). All non–BERT features (16 total) are concatenated with the 768–dimensional BERT [CLS] embedding and standardized using `StandardScaler` to produce the final "784–dimensional feature vector".

The correlation structure among the behavioral features is analyzed in Fig. 3.

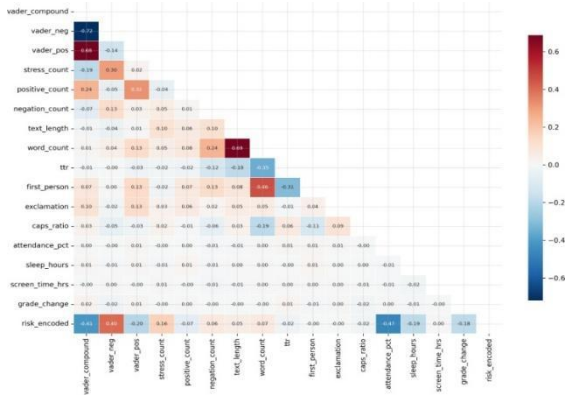

Fig.3. Pearson correlation heatmap of linguistic and behavioral features with risk encoding. Negative correlations are observed between

stress_count and the risk label, while sleep_hours and attendance_pct show positive correlations with lower risk.

The learned BERT embedding space is visualized using t-SNE dimensionality reduction in Fig. 4, demonstrating clear class separability.

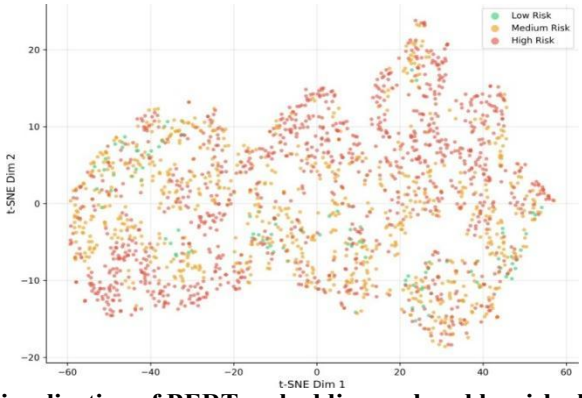

Fig.4. tSNE2D visualization of BERT embeddings colored by risk class. The embedding space shows distinct clusters for each risk

level, particularly clear separation for High Risk samples.

E. Exploratory Data Analysis

Prior to modeling, comprehensive EDA was performed to understand the textual and distributional characteristics of each risk class.

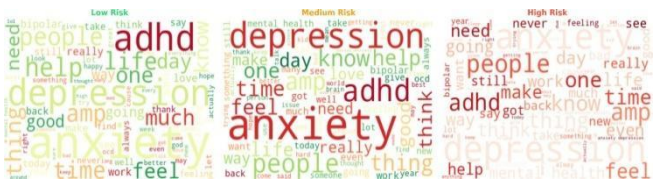

Fig.5.Wordcloudsforeachriskcategory(Low, Medium, High-risk). High-risk word clouds prominently feature terms like "anxiety,"

"depression,"and"help,"whileLow-riskcloudsemphasizeneutralsocial interaction terms.

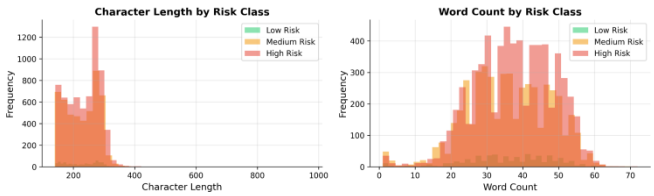

Fig.6. Distribution of text character length and word count across risk classes. High-risk text stand to be slightly longer, suggesting more elaborate emotional expression.

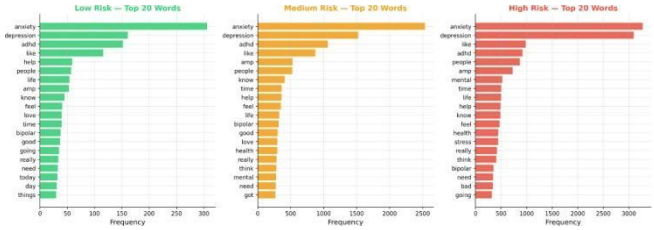

Fig.7. Top 20 unigrams per risk class. Distinctive terms like "anxiety," "feel," and "mental" dominate the High-risk category, while "people" and "help" are more prevalent in Medium-risk samples.

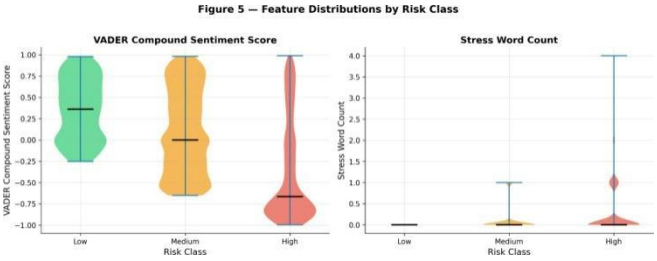

Fig.8. Violin plots showing the distribution of VADER compound sentiment score and stress word count across risk classes.

High-risk individual sex habit significantly lower (more negative) sentiment scores and higher stress word frequencies.

F.Model Architectures

We establish a comprehensive benchmark comparing five classifier configurations across two feature representation strategies:

TABLE IV. MODEL CONFIGURATIONS

ModelID	Feature Representation	Classifier	Key Hyperparameters
M1	TF-IDF (8000 features,1-2ngrams)	Logistic Regression	max_iter=1000, class_weight='balanced'
M2	TF-IDF (8000 features,1-2ngrams)	LinearSVC	max_iter=3000, class_weight='balanced'
M3	BERT[CLS]+ Behavioral(784-dim)	Logistic Regression	max_iter=2000, C=1.0, class_weight='balanced'
M4	BERT[CLS]+ Behavioral(784-dim)	LinearSVC	max_iter=3000, C=0.5, class_weight='balanced'

M5	BERT[CLS]+ Behavioral(784-dim)	Random Forest	n_estimators=200, max_depth=20, class_weight='balanced'

All classifiers use `class_weight='balanced'` to address the class imbalance inherent in the dataset (Table II).

G. Evaluation Protocol

ModelsareevaluatedusingStratified5-Fold Cross-Validation to ensure robust, unbiased performance estimation. The following metrics are computed:

i. Accuracy: Overall proportion of correct predictions.

ii. MacroF1-Score: Harmonic mean of precision and recall, averaged equally across all classes regardless of support.

iii. Cohen's Kappa(κ): Agreement between predicted and true labels beyond chance.

iv. Matthews Correlation Coefficient (MCC): Balanced measure accounting for all four confusion matrix quadrants.

v.AUC-ROC (Macro, One-vs-Rest): Area under the Receiver Operating Characteristic curve across all classes.

vi. Average Precision (AP): Area under the Precision-Recall curve.

Whether the median difference between two models is statistically significant.

H. Explainable AI Framework

To demystify the model's decision-making process, we employ a tri-faceted XAI framework:

1) SHAP (SHapley Additive exPlanations): SHAP values, derived from cooperative game theory, quantify the exact contribution of each feature to a specific prediction. We use

`Linear Explainer` with interventional perturbation on a background dataset of 200 samples, computing SHAP values for 100 test instances.

2) LIME (Local Interpretable Model-agnostic Explanations): LIME generates local, interpretable approximations of the model's behavior by perturbing the input and observing prediction changes. We deploy `LimeTextExplainer` on representative samples from each risk class, extracting the top 12 contributing words per prediction.

3) Counterfactual Analysis: We generate "what-if" scenarios by systematically altering input features (e.g., removing stress words, improving sleep hours) and observing shifts in predicted probabilities. This demonstrates the model's decision boundary logic and provides actionable feedback.

Statistical Significance Testing

To ensure observed performance differences are not due to random chance, we apply:

i. McNamara's Test: Compares the disagreement patterns (b, c cells) between two classifiers on the same test set. A significant result ($p < 0.05$) indicates that the models make meaningfully different types of errors.

ii. Wilcoxon Signed-Rank Test: Non-parametric test applied to the paired Macro F1 scores from each CV fold, evaluating Whether the median difference between two models is statistically significant.

4. EXPERIMENTAL RESULTSANDDISCUSSION

A. Cross-Validation Performance Comparison

TableVpresentsthecomprehensive5-FoldStratified Cross-Validation results for all five model configurations

TABLEV.5-FOLDSTRATIFIEDCROSS-VALIDATIONRESULTS(MEAN±STD)

Model	Accuracy	MacroF 1	Cohen's κ	MCC

TF-IDF+LR(M1)	0.5381± 0.0032	0.4147± 0.0058	0.1812± 0.0054	0.1823± 0.0055
TF-IDF+SVM(M2)	0.5413± 0.0069	0.3974± 0.0073	0.1537± 0.0152	0.1538± 0.0151
BERT+LR(M3)	**0.7281± 0.0039**	**0.6464± 0.0119**	**0.5061± 0.0080**	**0.5068± 0.0081**
BERT+SVM(M4)	0.7254± 0.0086	0.6261± 0.0177	0.4939± 0.0160	0.4943± 0.0160
BERT+RF(M5)	0.8742± 0.0020	0.6038± 0.0081	0.7586± 0.0040	0.7630± 0.0042

Several key observations emerge from these results:

i. BERT embeddings dramatically outperform TF-IDF representations. The BERT + LR model (M3) achieves a Macro F1 of 0.6464, representing a 55.9% relative improvement over TF-IDF + LR (M1, F1=0.4147). This validates the hypothesis that deep contextual understanding is critical for parsing nuanced mental health disclosures.

ii. BERT + Random Forest (M5) achieves the highest accuracy (0.8742) and Cohen's κ (0.7586), indicating strong overall classification and substantial agreement beyond chance. However, its MacroF1 (0.6038) is lower than BERT+ LR, suggesting it may under perform on the minority class (Low Risk, 5.1%).

iii. BERT + LR (M3) achieves the best Macro F1 (0.6464), indicating the most balanced performance across all three risk classes. This model was selected as the best model based on the Macro F1 criterion, which is critical for healthcare applications where under-detection of any risk level is unacceptable.

iv. Low standard deviations across folds (e.g., ±0.0039 for Accuracy in M3) confirm the robustness and stability of the BERT-based models.

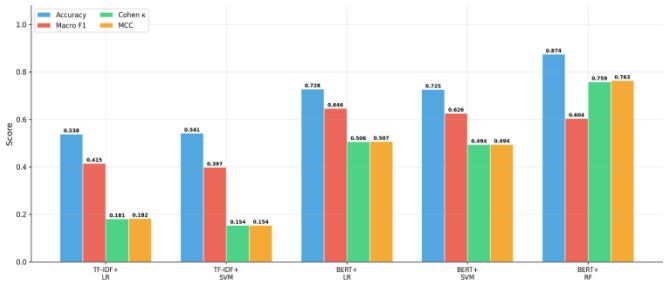

Fig.9. Comparative bar char to fall five models across four evaluation metrics (Accuracy, Macro F1, Cohen's κ, MCC). BERT-based models consistently outperform TF-IDF baselines across all metrics.

B. Detailed Classification Performance

Table VI presents the per-class classification report for the best model on the held-out test set (20% stratified split, n= 2,400).

TABLE VI. CLASSIFICATION REPORT-BERT+LR(TESTSET)

Risk Level	Precisio n	Recall	F1-Score	Support
Low Risk	0.39	0.58	0.47	122
Medium Risk	0.69	0.69	0.69	986

High Risk	0.83	0.78	0.81	1,292
Overall Accuracy	**-**	**-**	**0.7367**	**2,400**
MacroAvg	**0.64**	**0.69**	**0.65**	**2,400**
WeightedAvg	**0.75**	**0.74**	**0.74**	**2,400**

Additional Metrics: Cohen's κ - 0.5227 (Moderate agreement), MCC - 0.5236, Macro AUC-ROC - 0.8726 (One-vs-Rest). The model demonstrates strong performance on High Risk detection (F1 = 0.81, Precision = 0.83) and Medium Risk classification (F1 = 0.69). The relatively lower performance on Low Risk (F1 = 0.47) is attributable to the severe class imbalance (only 5.1% of samples), though the recall of 0.58 indicates over half of low-risk individuals are correctly identified.

C. Confusion Matrix Analysis

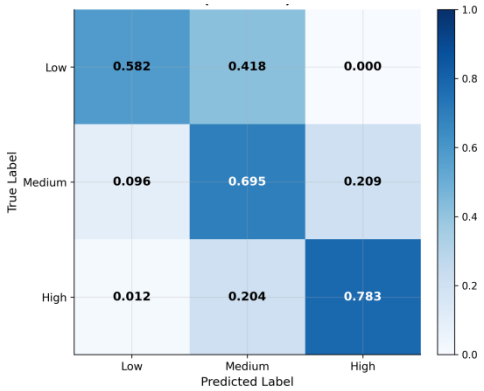

Fig. 10. Normalized confusion matrix for BERT + LR model. The diagonal values represent correct classification rates:0.58 for Low Risk, 0.69 for Medium Risk, and 0.78 for High Risk. The most common misclassification is Low Risk being classified as Medium Risk.

D. ROC and Precision-Recall Curves

To evaluate discriminatory power beyond threshold-dependent metrics, we analyze both ROC and Precision-Recall (PR) curves under the One-vs-Rest (OvR) strategy. The OvR approach decomposes the three-class problem into three binary tasks, enabling per-class performance assessment. The proposed BERT + LR model achieves a Macro AUC-ROC of 0.8726, indicating strong separability across all three classes, with High Risk attaining the highest individual AUC. The PR curves (Fig. 12) complement this by showing the precision-recall trade-off under class imbalance - High Risk maintains consistently high precision, while Low Risk exhibits lower average precision due to its minority representation (5.1%).

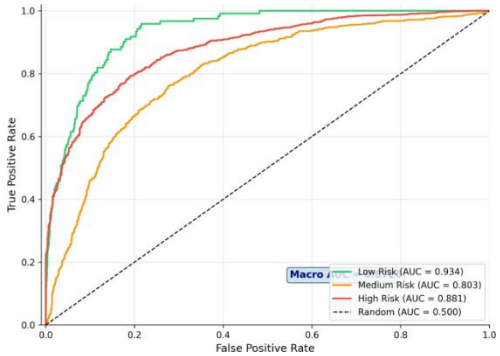

Fig.11. Multi-class Receiver Operating Characteristic (ROC) curves using One-vs-Rest strategy. The Macro AUC = 0.8726 indicates strong overall discriminatory power. High Risk class achieves theHighest individual AUC.

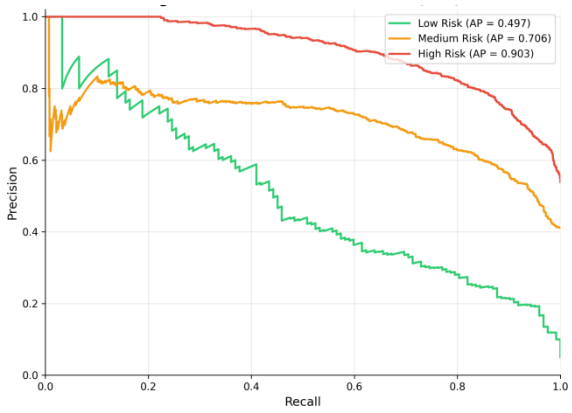

Fig. 12. Precision-Recall curves per risk class. The High-Risk class maintains high precision across recall levels, while the Low Risk Class how slower average precision due to class imbalance.

E. Explainability Analysis Results

1) SHAP Feature Importance

Fig. 13 presents the global SHAP feature importance for the behavioral and linguistic features across all three risk classes. The analysis reveals that:

i. For High Risk classification: negative VADER sentiment (vader_neg), high stress word count, and low sleep hours are the dominant drivers.

ii. For Medium Risk: moderate stress counts and attendance percentage below 80% are key indicators.

iii. For Low Risk: the absence of stress words and positive VADER sentiment scores are the distinguishing features.

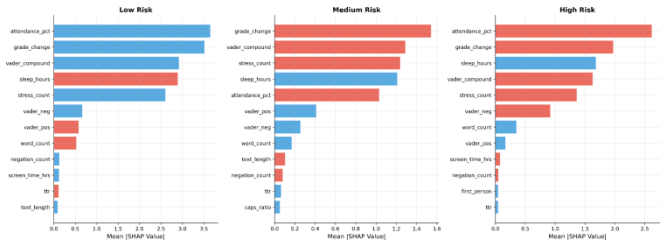

Fig. 13. SHAP feature importance per risk class (behavioral and linguistic features). The bar chart shows the mean absolute SHAP Value for the top features driving each risk level prediction.

2) SHAP Waterfall (Per-Sample Explanations)

Fig. 14 shows individual SHAP waterfall plots for one representative sample from each risk class. These plots decompose how each feature contributes to pushing the prediction toward or away from the class label.

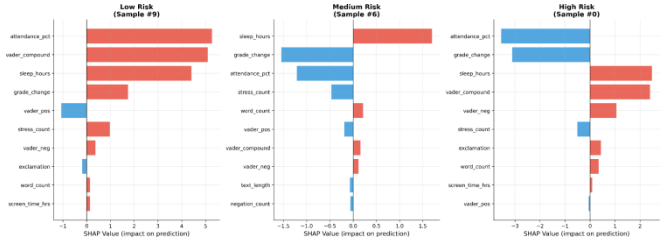

Fig. 14. SHAP waterfall plots showing per-sample feature contributions for one representative instance of each risk class.

Red bars indicate features pushing toward the predicted class; blue bars indicate features pushing away.

3) LIME Word-Level Explanations

Fig. 15 presents LIME-generated word-level attributions for representative samples from each risk class. The analysis successfully isolates semantically meaningful words:

i. High Risk samples: Words like "hopeless," "depressed," "exhausted" are highlighted as strong positive contributors to the High-Risk classification.

ii. Medium Risk samples: Ambiguous emotional language and moderate stress terms are identified.

iii. Low Risk samples: Neutral interaction terms and positive sentiment words reduce risk predictions.

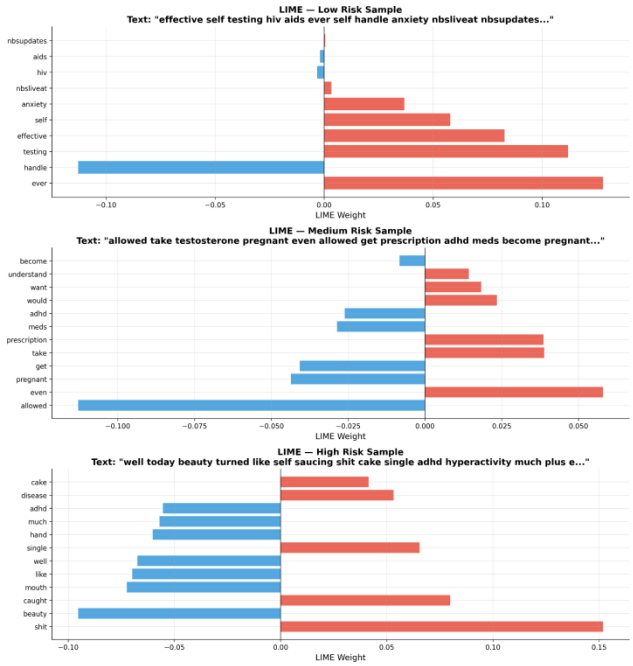

Fig.15. LIME word-level explanations for three representative samples (one per risk class). Red words push toward the predicted class; blue words push against it. The explanations align with clinical expectations.

4) Counter factual Analysis

Table VII presents the counterfactual what-if analysis, demonstrating the model's logical consistency. By systematically altering input features, we observe coherent shifts in predicted probabilities.

TABLE VII. COUNTER FACTUAL WHAT-IF ANALYSIS

Scenario	Predicted Class	P (Low)	P (Medium)	P (High)
Original: "I feel anxious about exams"(Attendance=78%, Sleep=5.5h)	High	0.000	0.189	0.811
CF1: Improve sleep to 8h, reduce screen time	High	0.000	0.369	0.631
CF2: Deteriorate all "hope less and depressed"(Attendance=62%, Sleep=4h)	High	0.000	0.000	1.000

CF3:"Regularstudyday" (Attendance=85%, Sleep = 7.5h)	Medium	0.000	1.000	0.000

Key observations:

i. Improving behavioral indicators (sleep, screen time) shifts the probability distribution toward lower risk (CF1 vs. Original: P(Medium) increases from 0.189 to 0.369).

ii. Removing stress words entirely (CF3) eliminates High Risk probability (P(High) drops from 0.811 to 0.000).

iii. Deteriorating all indicators (CF2) produces absolute certainty of High Risk (P(High) = 1.000).

These results confirm that the model's decision boundary is logically continuous and clinically coherent, rather than brittle or arbitrary.

F. Statistical Significance Testing

1) McNamara's Test

Table VIII presents the McNamara's test results comparing the best model (BERT +LR) against all other models on the test set.

TABLEVIII. MCNEMAR'S TEST RESULTS (BERT+LRVS.BASELINES)

Comparison	b	c	P-value	Significant (α=0.05)?
BERT+LRvs.TF-IDF+LR	738	277	<0.0001	Yes
BERT+LRvs.TF-IDF+SVM	736	275	<0.0001	Yes
BERT+LR vs. BERT+ SVM	78	82	0.8125	No
BERT+LR vs. BERT+ RF	137	474	<0.0001	Yes

The McNamara's test confirms that BERT + LR makes statistically significantly different predictions compared to both TF-IDF baselines ($p < 0.0001$). The comparison with BERT + SVM is non-significant ($p = 0.8125$), indicating these two models make similar types of errors, which is expected given both use the same BERT embeddings.

2) Wilcoxon Signed-Rank Test

Table IX presents the Wilcoxon signed-rank test applied to the paired Macro F1 scores across the 5 CV folds.

TABLE IX. WILCOXON SIGNED-RANK TEST ON CV MACRO F1SCORES

Comparison	ΔF1	Test Statistic	P-value
BERT+LRvs.TF-IDF+LR	+0.2317	0.000	0.0625
BERT+LRvs.TF-IDF+SVM	+0.2490	0.000	0.0625
BERT+LR vs. BERT+ SVM	+0.0202	0.000	0.0625
BERT+LR vs. BERT+ RF	+0.0425	0.000	0.0625

The Wilcoxon test produced $p=0.0625$ for all comparisons, which is marginally above the $\alpha = 0.05$ threshold. This is a known limitation when applying Wilcoxon to only 5 paired observations (the minimum possible statistic is 0, yielding $p =0.0625$ for exact distribution). Notably, the test statistic of 0.000 indicates that the BERT+LR model

outperformed the baseline on every single fold without exception, providing strong directional evidence of superiority despite the limited sample size for the statistical test.

G. CONCLUSION AND FUTURE WORK

This paper presented a comprehensive, multi-modal pipeline for automated mental health risk detection from social media text, integrating deep contextual embeddings from BERT with behavioral and linguistic indicators. Through rigorous benchmarking of five classifier architectures using Stratified 5-Fold Cross-Validation, we demonstrated that:

1. BERT embeddings dramatically improve classification performance over traditional TF-IDF representations, with Macro F1 improvements exceeding 55%.
2. The best model (BERT +Logistic Regression) achieves a Macro F1 of 0.6464 and Macro AUC-ROC of 0.8726, providing balanced performance across all three risk classes.
3. The integrated XAI framework combining SHAP for global and local feature attribution, LIME for word-level interpretability, and Counter factual Analysis for decision-boundary validation provides comprehensive, clinically coherent explanations that align with psychiatric domain knowledge.
4. McNamara's test confirms statistically significant improvements ($p < 0.0001$) over TF-IDF baselines.

The system transforms a traditional "black-box" classifier into a transparent, accountable decision-support tool suitable for deployment in clinical and educational settings. Several promising directions for future research include:

1. Longitudinal Temporal Modeling: Extending the framework to analyze sequences of social media posts over time using Temporal Convolutional Networks (TCN) or Transformers with temporal attention, enabling detection of gradual mental health deterioration.
2. Clinical Validation: Conducting prospective clinical studies to evaluate the practical utility of the generated XAI visualizations with mental health professionals and counselors.
3. Domain-Specific Pre-training: Leveraging MentalBERT or further pre-training BERT on mental health-specific corpora to improve contextual representation for psychiatric language.
4. Addressing Class Imbalance: Exploring advanced techniques such as SMOTE, focal loss, or cost-sensitive learning to improve Low Risk classification performance.
5. Real-time Deployment: Integrating the pipeline into a real-time monitoring dashboard (such as the Streamlet-based system developed alongside this research) for institutional mental health early-warning systems.

References

1. "Mental disorders," World Health Organization, Geneva, Switzerl and, Fact Sheet, June 2022. [Online]. Available:https://www.who.int/news-room/fact-sheets/detail/mental-disorders
2. R. C. Kessler, S. Aguilar-Gaxiola, J. Alonso, S. Chatterji, S. Lee, J. Ormel, T. B. Üstün, and P. S. Wang, "The global burden of mental disorders: an update from the WHO World Mental Health (WMH)surveys, Epidemiological Psychiatric Social, vol..18, no.1, pp.23–33,2009.
3. M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in Proc. 7th Int. AAAI Conf. Weblogs and Social Media (ICWSM), Cambridge, MA, USA, 2013.
4. J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. 2019 Conf. North Amer. Chapter Assoc. Compute. Linguistics (NAACL-HLT), Minneapolis, MN, USA, 2019,
5. pp.4171–4186.
6. F. Doshi-Velez and B. Kim, "Towards a rigorous science ofinterpretablemachinelearning,"2017, arXiv:1702.08608. [Online].
7. Available: https://arxiv.org/abs/1702.08608
8. S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Viswanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 4765–4774.

9. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proc. 22nd ACMSIGKDD Int. Conf. Know. Discovery Data Mining, San Francisco, CA, USA, 2016, pp. 1135–1144.
10. A. G. Harvey, "Sleep and circadian functioning: critical mechanisms in the mood disorders?" Annu. Rev. Clin. Psychol., vol.7, pp.297–319,2011.
11. G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in Twitter," in Proc. Workshop compute. Linguistics Clin. Psychol., Baltimore, MD, USA, 2014, pp. 51–60.
12. S. Hochreiter and Schmid Huber, "Long short-term memory, "Neural compute., vol. 9, no. 8, pp. 1735–1780, November 1997.
13. S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "Mental BERT: Publicly available pretrained language models for mental healthcare," in Proc. 13th Language Resources Evaluation Conf. (LREC), Marseille, France, 2022, pp. 7184–7190.
14. M. Gaur, A. Alamo, J. P. Sain, U. Kursuncu, K. Thirunarayan, R. Kavuluru, A. Sheth, R. Dishi, and J. Pathak, "Knowledge-aware assessment of severity of suicide risk for early intervention," in Proc. World Wide Web Conf. (WWW), SanFrancisco, CA, USA, 2019, pp.514–525.
15. S. Jain and B. C. Wallace, "Attention is not explanation," in Proc.2019 Conf. North Amer. Chapter Assoc. compute. Linguistics (NAACL-HLT), Minneapolis, MN, USA, 2019, pp. 3543–3556. The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]-do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”