

Mental Health Prediction Using Machine Learning and Explainable AI: A Review of Text-Based and Emotion-Aware Approaches

Santosh A. Chaudhri¹, Prof. (Dr.) Hemant H. Patel²

¹PG Scholar, Department Of Computer Science & Engineering, SOET, Dr. Subhash University, Junagadh, Gujarat, India

²Professor & Head Of Department, Department Of Computer Science & Engineering, SOET, Dr. Subhash University, Junagadh, Gujarat, India

Abstract - Mental health disorders have become an important global concern, while timely identification of individuals experiencing psychological difficulties remains challenging. The increasing use of social media has generated large volumes of user-generated text containing emotions, opinions, experiences, and behavioural signals that can support early mental health screening. Consequently, Natural Language Processing (NLP), Machine Learning (ML), Deep Learning (DL), and Transformer-based models have been increasingly investigated for mental health prediction from textual data. This review examines existing approaches for detecting mental health conditions from social media and other text-based sources, with emphasis on traditional machine learning, deep learning, and Transformer-based models. The reviewed studies address conditions including depression, anxiety, stress, suicidal ideation, and other mental health-related states. The review also investigates auxiliary information, particularly emotion-related features, to complement textual representations. In addition, Explainable Artificial Intelligence (XAI) techniques are examined to improve the transparency and interpretability of prediction models. The analysis identifies three major research gaps: insufficient cross-model benchmarking, limited integration of auxiliary emotion features, and limited explainability of prediction models. These gaps motivate the development of a comparative and explainable mental health prediction framework that evaluates different ML and Transformer-based approaches, investigates the contribution of emotion features, and provides interpretable explanations. The proposed framework is intended to support early screening and decision-making rather than replace professional mental health assessment.

Key Words: Mental Health Prediction, Machine Learning, Transformer Models, Emotion Features, SHAP, LIME.

1. INTRODUCTION

Mental health disorders such as depression, anxiety, and suicidal ideation have become important public-health concerns, creating a need for timely and scalable methods for identifying individuals who may require further support. Traditional mental-health assessment generally depends on clinical interviews, questionnaires, and expert evaluation, which can be time-consuming and may not always be suitable for continuous or large-scale screening. With the rapid growth of social media,

users increasingly express their thoughts, emotions, experiences, and psychological difficulties through text-based platforms. Consequently, social-media data has emerged as an important source for computational mental-health research and early detection of mental-health conditions. Natural Language Processing (NLP) and machine-learning techniques have been widely investigated for extracting mental-health-related signals from user-generated content. Early approaches primarily relied on manually designed linguistic features, statistical representations, and conventional machine-learning algorithms. Tadesse et al. [1], for example, demonstrated that NLP and machine-learning methods could be used to identify depression-related posts on Reddit. Similarly, Haque et al. [2] compared NLP, machine-learning, and deep-learning approaches for suicidal-ideation detection and reported that deep-learning models could effectively capture signals associated with suicidal thoughts. Nadeem et al. [5] further developed a hybrid Sequence, Semantic, Context Learning (SSCL) framework using CNN, LSTM, GRU, and self-attention mechanisms, demonstrating the potential of deep-learning approaches for detecting depression from social-media text. The development of deep neural networks has subsequently improved the ability of computational models to capture semantic and contextual information. However, conventional recurrent and convolutional architectures can face difficulties in modeling long-range dependencies and complex contextual relationships in social-media language. Transformer architectures have therefore attracted considerable attention because of their ability to learn contextual representations from text. Bokolo and Liu [8] conducted a comparative evaluation of machine-learning and Transformer-based techniques for depression and suicide detection and reported that Transformer models such as RoBERTa and DeBERTa provided stronger performance than traditional approaches. Their subsequent comparative study also identified RoBERTa as a highly effective model for depression detection from Twitter data [9]. BERT-based architectures have further enabled researchers to develop models specifically suited to mental-health-related language. Abbas et al. [10] proposed a Transformer-based contextualized embedding approach combined

with probabilistic features for depression detection from social-media content. Sao and Lim [11] developed MIROBERTa using transfer learning on mental-health-related Reddit subreddits, demonstrating the usefulness of domain-specific transfer learning for mental-illness text classification. More recently, Hasan et al. [13] compared several Transformer architectures, including BERT, RoBERTa, DistilBERT, ALBERT, and ELECTRA, against LSTM-based models for multi-class mental-health classification. Their findings indicate that Transformer-based models generally outperform conventional alternatives across multiple mental-health categories. An important direction in recent research is the movement from simple binary depression detection toward broader and more detailed mental-health assessment. Most early systems concentrated on determining whether depression or suicidal ideation was present. However, real-world mental-health conditions are more complex and may involve multiple disorders and different severity levels. Hasan et al. [13] investigated multi-class classification involving several mental-health conditions, while Hussain et al. [14] focused on multi-level depression-severity detection using deep Transformer and machine-learning techniques. These studies demonstrate the growing importance of models that can distinguish between different mental-health conditions and estimate their severity rather than providing only a binary prediction. Another important development is the incorporation of additional linguistic and contextual information. Social-media text contains informal language, emotional expressions, slang, abbreviations, and implicit meanings that may not be adequately represented by conventional text features alone. Wawale et al. [15] investigated a Hybrid Mental Health Detection Framework that combines domain-specific BERT representations with psycholinguistic features derived from LIWC. Such hybrid approaches demonstrate the potential benefit of combining contextual semantic representations with structured linguistic information. Similarly, Baydili et al. [4] investigated feature-selection techniques together with machine-learning and pretrained language-model representations for depression and suicidal-tendency detection across multiple datasets.

Explainability is another critical requirement for mental-health prediction systems. Although Transformer and deep-learning models can achieve high predictive performance, their decision-making processes may be difficult to interpret. In mental-health applications, understanding why a model produces a particular prediction is especially important because incorrect predictions can have serious consequences. Xin and Zakaria [16] addressed this issue by developing BERT-BiLSTM and BERT-CNN models and examining their decision-making using Transformer interpretability and attention-based visualization techniques. Their work demonstrates that explainability can provide insights into the linguistic patterns that influence model

predictions while maintaining the advantages of Transformer-based representations.

Despite the progress achieved by these approaches, several challenges remain. Existing studies often evaluate different model families under different datasets, preprocessing strategies, and experimental settings, making direct comparison of their performance difficult. In addition, many studies primarily rely on textual representations without systematically examining the contribution of auxiliary features such as emotional information. Furthermore, although deep-learning and Transformer-based models can achieve high predictive performance, their decision-making processes may remain difficult to interpret. These challenges highlight the need for systematic cross-model benchmarking, integration of relevant auxiliary features, and explainable prediction methods for mental-health text classification.

Overall, existing research demonstrates a clear progression from conventional NLP and machine-learning methods toward deep-learning and Transformer-based architectures for mental-health detection from textual data. However, the literature also indicates that high predictive performance alone is insufficient for practical deployment. Systematic comparison across model families, incorporation of informative auxiliary features, explainability, multi-class and severity prediction, linguistic diversity, privacy, fairness, and responsible use must also be considered. Therefore, this review paper examines the evolution of machine-learning, deep-learning, and Transformer-based approaches for mental-health prediction from text, with particular emphasis on cross-model benchmarking, explainable artificial intelligence (XAI), auxiliary feature integration, multi-class and severity detection, and the remaining research challenges.

2. LITERATURE REVIEW

Research on automated mental health detection from social media has progressed from traditional machine learning and manually engineered linguistic features toward deep learning and Transformer-based architectures. The literature reviewed in this study demonstrates a gradual shift from detecting individual disorders, particularly depression and suicidal ideation, toward multi-class classification, severity estimation, domain-specific language modeling, explainability, and cross-dataset benchmarking.



FIG. 1. EVOLUTION OF RESEARCH METHODS IN MENTAL HEALTH PREDICTION (2019–2026)

2.1 TRADITIONAL MACHINE LEARNING AND LINGUISTIC FEATURE-BASED APPROACHES

Early research on social-media-based mental health detection primarily relied on Natural Language Processing (NLP), manually engineered linguistic features, and conventional machine-learning algorithms. Tadesse et al. [1] investigated depression-related posts on Reddit using linguistic features including LIWC, Latent Dirichlet Allocation (LDA), and bigrams. Their experiments demonstrated that combining multiple feature types improved classification performance, with the combination of LIWC, LDA, and bigram features achieving an F1-score of 0.93 using a Multilayer Perceptron classifier. This study highlighted the importance of linguistic and psycholinguistic characteristics in identifying depression-related language.

Haque et al. [2] conducted a comparative analysis of machine-learning and deep-learning methods for suicidal-ideation detection using Twitter data. Their study used traditional feature extraction techniques such as CountVectorizer and word embeddings and compared several classifiers. Random Forest achieved the strongest performance among the traditional machine-learning models, while a BiLSTM model using word embeddings achieved an accuracy of 93.6% and an F1-score of 0.93. These findings indicate that although conventional machine-learning methods can provide competitive results, representation learning can improve the ability to capture patterns associated with suicidal ideation.

More recently, Sood et al. [3] evaluated eight machine-learning algorithms for detecting anxiety and depression from Twitter data, including Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Random Forest, Naive Bayes, Decision Tree, Gradient Boosting, and AdaBoost. Their results varied across datasets, with Random Forest performing strongly on one dataset and Logistic Regression and SVM providing competitive results on another. This

variation demonstrates that the effectiveness of a machine-learning algorithm can depend strongly on the dataset and task.

Baydili et al. [4] further investigated machine-learning-based detection of depression and suicidal tendencies using six datasets collected from Twitter and Reddit. Their framework combined pretrained language-model representations with feature-selection techniques and SVM classification. Reported accuracy varied considerably across datasets, ranging from 80.74% to 99.96%. This variation emphasizes the importance of dataset characteristics and experimental settings when comparing mental-health detection systems.

Overall, traditional machine-learning approaches demonstrate that linguistic features can provide useful signals for mental-health detection. However, their dependence on manually designed features can limit their ability to capture complex contextual relationships, implicit meanings, and variations in informal social-media language.

2.2 DEEP LEARNING APPROACHES

Deep-learning approaches were introduced to overcome some of the limitations associated with manually engineered features by learning representations directly from text. Nadeem et al. [5] proposed a hybrid Sequence, Semantic, Context Learning (SSCL) framework using CNN, LSTM, GRU, and self-attention mechanisms for depression detection. The framework achieved an accuracy of 97.4% for binary classification and 82.9% for ternary classification. The authors also reported an F1-score of 94.4% when testing on unseen tweets. The study demonstrates the value of combining sequence modeling, semantic representations, and contextual attention for depression detection.

Haque et al. [2] similarly demonstrated the effectiveness of BiLSTM models with word embeddings for suicidal-ideation detection. These results suggest that recurrent neural networks can capture sequential relationships in social-media text that may not be adequately represented by conventional feature-based classifiers.

Zhou and Mohd [6] developed a framework combining BERT contextual embeddings with BiLSTM to address the informal nature of social-media text. Their preprocessing pipeline included emoji normalization, slang replacement, and emotion scores, while the BERT-BiLSTM architecture was used to represent contextual and sequential information. This approach specifically addresses a major characteristic of social-media text: users frequently communicate using slang, emojis, abbreviations, and context-dependent expressions.

Arshad et al. [7] investigated depression detection across multiple social-media platforms using traditional machine learning and deep-learning approaches. Their study incorporated text normalization, tokenization,

lemmatization, and random oversampling to address noise and class imbalance. A GRU-Attention model demonstrated strong contextual modeling capabilities, while an ensemble of successful models achieved a reported accuracy of 94.74%. The multi-platform design is particularly relevant because models trained on a single platform may not generalize effectively to other social-media environments.

Although deep-learning models provide improved representation learning, their performance may still depend on the quality and quantity of training data and on the ability of the architecture to model long-range contextual relationships. These limitations contributed to the growing adoption of Transformer-based models.

2.3 TRANSFORMER-BASED APPROACHES

Transformer architectures have become an important direction in computational mental-health research because they provide contextualized representations and can model relationships between words over longer textual contexts. Many reviewed studies report stronger performance from Transformer-based models than traditional machine-learning approaches; however, performance varies across datasets, tasks, and evaluation settings. Their results generally favored Transformer architectures, particularly RoBERTa and DeBERTa, although traditional Logistic Regression performed better on one dataset. This finding is important because it indicates that Transformer models are not universally superior and that dataset characteristics can influence model performance.

In another study, Bokolo and Liu [9] evaluated depression detection using a dataset containing 632,000 tweets. The investigated models included Logistic Regression, Bernoulli Naive Bayes, Random Forest, DistilBERT, SqueezeBERT, DeBERTa, and RoBERTa. RoBERTa achieved the highest reported accuracy of 98.1%, with a mean accuracy of approximately 0.97 across ten cross-validation folds. The large dataset provides evidence of the effectiveness of Transformer-based representations for large-scale social-media depression detection.

Abbas et al. [10] proposed a BERT-RF feature-engineering approach in which BERT was used to extract contextualized embeddings and Random Forest was used to generate probabilistic features. These features were subsequently used by several classifiers. The proposed approach achieved a reported accuracy of 99% using Logistic Regression. This study demonstrates that Transformer representations can also be combined with traditional classifiers rather than being used only as end-to-end classifiers.

Sao and Lim [11] investigated domain adaptation through MIROBERTa, a model pretrained on mental-health-related Reddit content. Their study covered seven mental illnesses and compared traditional machine-learning models, deep-learning architectures, and Transformer models. Domain adaptation improved the

performance of the Transformer representation, with MIROBERTa achieving an accuracy and F1-score of 0.847. The findings indicate that pretraining on domain-specific mental-health language can improve contextual representation.

Dehkharghani [12] investigated Transformer-based methods for depression and suicidal-ideation detection across the SuicideDetection, CEASEv2.0, and SWMH datasets. The proposed approach achieved F1-scores of 0.97 and 0.75 on two datasets, while performance was lower on the SWMH dataset. The variation across datasets again highlights the importance of external validation and demonstrates that strong performance on one dataset does not necessarily guarantee equivalent performance elsewhere.

2.4 MULTI-CLASS MENTAL HEALTH CLASSIFICATION

A significant limitation of earlier research is its concentration on individual disorders or binary classification. More recent studies have expanded the problem toward identifying multiple mental-health conditions.

Hasan et al. [13] performed a comparative evaluation of BERT, RoBERTa, DistilBERT, ALBERT, and ELECTRA against LSTM-based architectures for multi-class mental-health classification. The dataset covered six mental-health conditions and a control category. Transformer models consistently outperformed the evaluated alternatives, with RoBERTa achieving F1-scores and accuracies in the 91–99% range across classes. However, attention-enhanced LSTM models using BERT embeddings also approached Transformer performance while requiring less training time. This demonstrates an important accuracy-efficiency trade-off for practical deployment.

Sao and Lim [11] similarly addressed multi-class classification involving seven common mental illnesses. Their results indicate that domain-adapted Transformer models can provide useful improvements over standard pretrained models. Together, these studies indicate a shift from simple depression detection toward broader mental-health classification.

2.5 DEPRESSION SEVERITY DETECTION

Beyond identifying whether depression is present, recent research has begun to investigate the severity of depressive symptoms. Hussain et al. [14] proposed a multi-level depression severity detection framework using Reddit posts. Four severity categories—minimum, mild, moderate, and severe—were considered. The study compared ten classical machine-learning classifiers with ten Transformer architectures. Among the traditional approaches, XGBoost achieved an F1-score of 94.01 using GloVe embeddings, whereas MentalBERT achieved the highest overall F1-score of 97.31 among the Transformer models.

This work is important because severity prediction provides more detailed information than binary

classification. However, the literature represented in the reviewed dataset indicates that severity detection remains less explored than binary mental-health classification. Further investigation is therefore required to determine whether severity models can generalize across different datasets, platforms, demographic groups, and annotation schemes.

2.6 AUXILIARY AND PSYCHOLINGUISTIC FEATURES

Although Transformer models provide contextual representations, several studies demonstrate the value of combining contextual embeddings with additional linguistic information. Tadesse et al. [1] showed that combining LIWC, LDA, and bigram features improved depression detection compared with individual feature types.

Wawale et al. [15] extended this idea by proposing a Hybrid Mental Health Detection Framework (HMHDF) that combines domain-specific BERT representations with a 73-dimensional LIWC 2022 linguistic feature vector. The features are fused before multi-label classification. The framework was evaluated using the CLPsych 2019 and RSDD datasets and reported an F1-score of 0.89 and an AUC of 0.93. The study also emphasizes ethical considerations including consent, demographic fairness, misclassification, and potential misuse.

Zhou and Mohd [6] incorporated emotion scores together with BERT and BiLSTM representations, while Abbas et al. [10] combined BERT contextual embeddings with probabilistic features generated through Random Forest. These studies collectively suggest that combining semantic representations with structured linguistic or statistical features can provide complementary information.

2.7 EXPLAINABLE AI AND MODEL INTERPRETABILITY

High-performing deep-learning and Transformer models can be difficult to interpret, creating an important challenge for mental-health applications. Explainability is particularly relevant because predictions concerning depression or suicidal ideation can have significant consequences.

Xin and Zakaria [16] investigated explainable depression detection using BERT, BERT-BiLSTM, and BERT-CNN and compared these models with MentalBERT. Their study used three publicly available datasets and examined model decision-making through Transformer interpretability and attention visualization. BERT-CNN achieved reported accuracies of 0.982, 0.961, and 1.0 across the three evaluated datasets. The attention visualizations provided information about the language patterns influencing model decisions.

Wawale et al. [15] also emphasize responsible AI considerations in their hybrid framework, including fairness, consent, and misclassification risk. These

studies indicate that model performance should not be considered independently of interpretability and responsible deployment.

Nevertheless, explainability remains an emerging area within mental-health prediction. More research is required to determine whether explanation methods provide faithful representations of model reasoning and whether explanations are understandable and useful to mental-health professionals.

2.8 OVERALL SYNTHESIS OF THE LITERATURE

The reviewed research demonstrates a clear progression in computational mental-health detection. Traditional machine-learning methods established the importance of linguistic and psycholinguistic features, while deep-learning architectures improved automated representation learning. Transformer-based models subsequently provided stronger contextual representations and have generally achieved high performance across several mental-health classification tasks.

However, the literature also demonstrates that performance varies substantially across datasets and tasks. Transformer models often outperform traditional approaches, but exceptions have been reported, particularly when datasets contain simpler patterns or when the evaluation setting differs. Furthermore, high performance on a single dataset does not necessarily indicate robust generalization.

Recent research is consequently moving toward domain-specific Transformers, multi-class classification, severity estimation, explainable AI, large-scale benchmark datasets, multimodal and auxiliary feature integration, and multilingual evaluation. These developments provide a foundation for identifying the remaining research gaps and defining directions for more robust, explainable, and generalizable mental-health prediction systems.

3. COMPARATIVE ANALYSIS OF EXISTING STUDIES

The reviewed studies demonstrate the progressive development of Natural Language Processing (NLP), machine learning, deep learning, and Transformer-based techniques for mental health detection from social media and conversational text. The studies differ in terms of datasets, mental health conditions, classification tasks, feature representations, model architectures, and evaluation metrics. Table 1 presents a comparative analysis of the 20 selected studies.

TABLE 1 : COMPARATIVE ANALYSIS OF EXISTING STUDIES

Ref .	Auth or/Y ear	Dataset	Task	Method	Key Finding
1	Tadesse et al. (2019)	Reddit	Depression detection	LIWC, LDA, bigram features with SVM/MLP	Demonstrated that linguistic, topic, and textual features can effectively identify depression-related Reddit posts.
2	Haque et al. (2022)	Twitter; 49,178 instances	Suicidal ideation detection	Random Forest, BiLSTM, word embeddings	Both traditional ML and deep learning were effective; BiLSTM achieved approximately 93% accuracy/F1 around 0.93.
3	Sood et al. (2025)	Twitter / DATD datasets	Anxiety and depression detection	LR, KNN, SVM, RF, NB, DT, GBM, AdaBoost	RF achieved about 92.6% accuracy on one dataset, while LR/SVM were

					competitive on another.
4	Baydili et al. (2025)	Twitter and Reddit; multiple datasets	Depression and suicidal tendency detection	CWINCA feature selection, SVM, pretrained language models	Demonstrated the usefulness of feature selection and machine-learning/pretrained representations across multiple datasets.
5	Nadeem et al. (2022)	Twitter	Depression detection	CNN, LSTM, GRU, self-attention, SSCL	Proposed a hybrid SSCL framework combining sequence, semantic, and contextual information; reported strong performance for binary and ternary classification.
6	Zhou & Mohd (2025)	Social-media text	Depression / mental-health detection	BERT-BiLSTM, emotion-related features, emoji/slang handling	Showed the potential of contextual modeling

					combined with informal - language and emotional information.
7	Arshad et al. (2026)	Twitter/X, Reddit, Facebook	Depression detection	ML/DL, GRU-Attention, ensemble methods	Demonstrated the use of contextual and feature-engineering approaches for depression detection across social-media platforms.
8	Bokolo & Liu (2024)	Sentiment140 and Suicide-Watch	Depression and suicide detection	LR, NB, RF, RoBERTa, DeBERTa, DistilBERT, SqueezeBERT	Showed that model performance varies across datasets, highlighting the importance of comparative benchmarking.
9	Bokolo & Liu (2023)	632,000 tweets	Depression detection	LR, NB, RF, DistilBERT, SqueezeBERT, DeBERTa, RoBERTa	RoBERTa reported approximately 98.1% accuracy, demonstrating

					strong performance on large-scale Twitter data.
10	Abbas et al. (2024)	20,000 tagged tweets	Depression detection	BERT contextual embeddings, probabilistic features, ML classifiers	Demonstrated that combining contextual BERT representations with additional probabilistic/ML features can improve depression classification.
11	Sao & Lim (2024)	Reddit; 54,161 posts from 11 subreddits	Multi-class mental-illness classification	BERT, RoBERTa, BigBird, Longformer, MIMRoBERTa	Demonstrated the usefulness of domain-specific transfer learning through MIMRoBERTa for mental-health text classification.

12	Deh khar ghan i (2024)	Suicide Detection, CEASE v2.0, SWMH	Mental-disorder and suicidal-ideation detection	Transformer-based deep learning	Demonstrated the applicability of Transformer models across multiple mental-health classification tasks and datasets.
13	Hasan et al. (2025)	Reddit	Multi-class mental-health classification	BERT, RoBERTa, DistilBERT, ALBERT, ELECTRA, LSTM	Demonstrated strong performance of Transformer architectures for multi-class mental-health classification.
14	Hussain et al. (2025)	Reddit	Multi-level depression-severity detection	BERT, RoBERTa, MentalBERT, XGBoost and other ML models	Mental BERT reported approximately 97.31 F1, demonstrating the potential of domain-specific Transformers for severity classification.
15	Wa	CLPsysc	Depre	BERT +	Demons

	wale et al. (2026)	h 2019, RSDD	ssion, anxiety and suicidal-ideation detection	LIWC	trated that combining contextual BERT representations with psycholinguistic features can support mental-health classification.
16	Xin & Zakaria (2024)	Depression Reddit, Mental Health Corpus, Tweets	Explorable depression detection	BERT, BERT-BiLSTM, BERT-CNN, MentalBERT; attention visualization	Demonstrated that combining Transformer representations with CNN/BiLSTM architectures can provide strong depression-detection performance while supporting interpretability.
17	Mohammed & Hassani (2025)	Sorani Kurdish Twitter; 960 posts	Depression detection	SVM, NB, LR, RF	RF achieved approximately 80% accuracy, demonstrating

					the potential of ML approaches for lower-resource language settings.
18	Sun et al. (2024)	Counseling and psychotherapy transcripts	Anxiety and depression classification	BERT, RoBERTa, Longformer, Mistral-7B, SVM, GPT-based models	Found that larger language models did not necessarily outperform traditional ML approaches on long counseling/psychotherapy transcripts..
19	Raihan et al. (2024)	Mental Health; millions of social-media instance	Multi-task mental-health detection	Semi-supervised labeling, Flan-T5, DisorBERT, MentalBERT	Introduced a large-scale multi-task resource for mental-health research and demonstrated the value of large-scale social-media data.
20	Hasan &	Multiple	Mental-	Multiple ML and	Provided a benchm

	Saquer (2026)	Reddit-derived datasets	health detection benchmarking	Transformer approaches	ark-oriented framework for more systematic evaluation of mental-health detection models.
--	---------------	-------------------------	-------------------------------	------------------------	--

Methods Used in Mental Health Prediction Research (20 Papers)

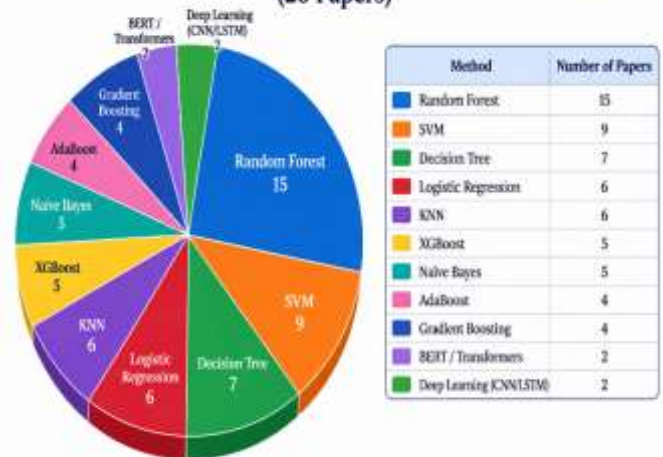


FIG. 2 : DISTRIBUTION OF METHODS USED ACROSS THE REVIEWED STUDIES

4. COMPARATIVE DISCUSSION

The comparative analysis shows that transformer-based models generally provide stronger performance than traditional machine learning and conventional deep learning methods for mental health detection. Models such as BERT, RoBERTa, and MentalBERT achieve high performance across depression, suicidal ideation, and multi-class mental health tasks. Hybrid approaches combining transformers with CNN, BiLSTM, attention, or linguistic features further improve detection performance and interpretability. However, the reported results vary considerably across datasets and tasks, indicating that model performance alone is not sufficient for reliable mental health prediction. Existing studies also demonstrate the importance of attention mechanisms and linguistic features for explaining model decisions, providing a basis for developing more interpretable transformer-based mental health prediction systems.

5. RESEARCH GAPS

The review of existing studies identifies three major research gaps in text-based mental health prediction.

5.1 CROSS-MODEL BENCHMARKING

Existing studies use different datasets, pre-processing methods, feature representations, and evaluation settings, making direct comparison between machine learning and Transformer models difficult [3], [4], [8], [9], [11], [12], [13], [18], [20]. A systematic comparison under a common experimental framework is therefore required to understand the relative performance of classical ML and Transformer-based models for multiclass mental health prediction.

5.2 AUXILIARY FEATURE INTEGRATION

Most text-based approaches primarily rely on textual features. However, emotional characteristics expressed in text may provide additional information for identifying mental health-related patterns. Previous studies have explored linguistic, psycholinguistic, probabilistic, and emotion-related features alongside textual representations [1], [6], [7], [10], [15]. However, limited work has systematically investigated the integration of emotion features with Transformer-based text representations for multiclass mental health prediction. This creates an opportunity to examine whether emotion-enhanced representations can improve multiclass prediction performance.

5.3 EXPLAINABILITY

Transformer and other deep learning models can provide strong predictive performance but are often difficult to interpret. In mental health applications, understanding why a model produces a particular prediction is important for transparency and responsible use [15], [16]. Therefore, there is a need to apply Explainable AI (XAI) techniques such as SHAP and LIME to identify and interpret the textual and auxiliary features contributing to model predictions.

6. PROPOSED RESEARCH FRAMEWORK

Based on the identified research gaps, this review proposes a framework for multiclass mental health prediction from textual data. The framework combines conventional machine learning, Transformer-based models, emotion features, and Explainable AI.

First, the 53K mental health dataset is preprocessed to clean and normalize the textual data, followed by a stratified train-validation-test split. Classical machine learning models such as Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF) are evaluated alongside BERT and RoBERTa under a common experimental setup for cross-model benchmarking.

Next, six emotion features are extracted from the text and integrated with textual representations through feature fusion. The resulting text-only and text-plus-emotion models are compared using an ablation analysis to determine the contribution of emotion information.



FIG. 3 : PROPOSED RESEARCH FRAMEWORK

Finally, Explainable AI techniques such as SHAP and LIME are applied to the selected models to identify the textual and emotion-related features contributing to predictions. The framework therefore addresses three key areas: model benchmarking, emotion feature integration, and explainability.

7. EVALUATION FRAMEWORK

The proposed framework will be evaluated using a common experimental setup to ensure a fair comparison among different models.

7.1 DATASET SPLIT

After preprocessing and removing missing or duplicate text records, the dataset is divided using stratified sampling so that the class distribution is maintained across the subsets. The data can be divided into training, validation, and testing sets. The training set is used for model learning, the validation set for model selection and parameter tuning, and the test set is used only for final performance evaluation.

7.2 EVALUATION METRICS

The models are evaluated using Accuracy, Macro Precision, Macro Recall, Macro-F1, and Weighted-F1. Since the dataset contains multiple classes with an imbalanced distribution, Macro-F1 is given particular importance because it gives equal weight to each class. Confusion matrices and per-class performance are also used to analyze classification errors.

7.3 COMPARISON PROCEDURE

All models are evaluated using the same dataset split and consistent preprocessing and evaluation criteria. LR, SVM, and RF are compared with BERT and RoBERTa to establish a cross-model benchmark. The results are compared using the selected evaluation metrics to identify differences in performance between classical machine learning and Transformer-based approaches.

7.4 ABLATION EVALUATION

An ablation study is conducted to measure the contribution of emotion features. Two configurations are compared:

- **Text-only model:** Uses textual features without emotion information.
- **Text + Emotion model:** Combines textual representations with six extracted emotion features.

The performance difference between these configurations is analyzed to determine whether integrating emotion features provides additional predictive value.

8. ETHICAL CONSIDERATIONS

Mental health prediction using textual data involves sensitive information and therefore requires careful and responsible use. Privacy and confidentiality should be maintained when collecting and processing textual data, and personally identifiable information should be protected.

Bias and fairness are also important considerations because differences in class distribution, data sources, and language patterns may affect model performance across different groups. Appropriate evaluation and transparent reporting are therefore necessary.

The predictions generated by machine learning and Transformer models should be considered as decision-support or preliminary screening information rather than

clinical diagnosis. Human and professional judgment should remain important when interpreting predictions.

Finally, Explainable AI (XAI) techniques such as SHAP and LIME can improve transparency by showing which features contribute to model predictions. However, explanations should be interpreted carefully and should not be considered a substitute for professional clinical assessment.

9. DISCUSSION

The reviewed studies show that both traditional machine learning and Transformer-based models are useful for mental health prediction, but differences in datasets and evaluation settings make direct comparison difficult [3], [8], [9], [13], [20]. Emotion and other auxiliary features can provide additional information beyond text and may improve classification when combined with contextual representations [1], [6], [10], [15]. Furthermore, XAI techniques such as attention visualization, SHAP, and LIME can improve the interpretability of complex models [15], [16]. Therefore, the proposed framework focuses on cross-model benchmarking, emotion feature integration, and explainability for multiclass mental health prediction.

10. CONCLUSION

This review examined machine learning and Transformer-based approaches for text-based mental health prediction. The literature highlights the importance of multiclass prediction, systematic cross-model benchmarking, emotion feature integration, and model explainability. Based on the identified gaps, a framework was proposed that compares classical ML and Transformer models, integrates emotion features with textual representations, and applies Explainable AI techniques such as SHAP and LIME. The framework provides a structured direction for developing more transparent and interpretable mental health prediction systems from textual data.

11. REFERENCES

- [1] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of depression-related posts in Reddit social media forum," *IEEE Access*, vol. 7, pp. 44883–44893, 2019, doi: 10.1109/ACCESS.2019.2909180.
- [2] R. Haque, N. Islam, M. Islam, and M. M. Ahsan, "A comparative analysis on suicidal ideation detection using NLP, machine, and deep learning," *Technologies*,

vol. 10, no. 3, Art. no. 57, 2022, doi: 10.3390/technologies10030057

[3] T. Sood, P. Garg, H. Pannu, and S. Dullat, "Machine learning algorithms and NLP for anxiety and depression detection using Twitter data," in Proc. 2025 Int. Conf. Cognitive Computing in Engineering, Communications Sciences and Biomedical Health Informatics (IC3ECSBHI), 2025, pp. 450–455, doi: 10.1109/IC3ECSBHI63591.2025.10990488.

[4] İ. Baydili, B. Tasci, and G. Tasci, "Deep learning-based detection of depression and suicidal tendencies in social media data with feature selection," Behavioral Sciences, vol. 15, no. 3, Art. no. 352, 2025, doi: 10.3390/bs15030352.

[5] A. Nadeem, M. Naveed, and M. I. Satti, "Depression detection based on hybrid deep learning SSCL framework using self-attention mechanism: An application to social networking data," Sensors, vol. 22, no. 24, Art. no. 9775, 2022, doi: 10.3390/s22249775.

[6] S. Zhou and M. Mohd, "Mental health safety and depression detection in social media text data: A classification approach based on a deep learning model," IEEE Access, vol. 13, pp. 63284–63297, 2025, doi: 10.1109/ACCESS.2025.3559170.

[7] I. Arshad, S. Naseem, M. Mussarat, A. Zanib, S. Al-Otaibi, A. Yousif, T. Mahmood, and A. R. Khan, "A multi-platform approach to optimizing depression detection in social media through feature engineering and contextual analysis," International Journal of Computational Intelligence Systems, vol. 19, Art. no. 97, 2026, doi: 10.1007/s44196-026-01225-y.

[8] B. G. Bokolo and Q. Liu, "Advanced comparative analysis of machine learning and Transformer models for depression and suicide detection in social media texts," Electronics, vol. 13, no. 20, Art. no. 3980, 2024, doi: 10.3390/electronics13203980.

[9] B. G. Bokolo and Q. Liu, "Deep learning-based depression detection from social media: Comparative evaluation of ML and Transformer techniques," Electronics, vol. 12, no. 21, Art. no. 4396, 2023, doi: 10.3390/electronics12214396.

[10] M. A. Abbas, K. Munir, A. Raza, N. Abdel Samee, M. M. Jamjoom, and Z. Ullah, "Novel Transformer based contextualized embedding and probabilistic features for depression detection from social media,"

IEEE Access, vol. 12, pp. 54087–54100, 2024, doi: 10.1109/ACCESS.2024.3387695.

[11] M. Sao and H.-J. Lim, "MIroBERTa: Mental illness text classification with transfer learning on subreddits," IEEE Access, vol. 12, pp. 197454–197466, 2024, doi: 10.1109/ACCESS.2024.3522465.

[12] R. Dehkharghani, "Mental disorder and suicidal ideation detection from social media using deep neural networks," Journal of Computational Social Science, vol. 7, 2024, doi: 10.1007/s42001-024-00307-1.

[13] K. Hasan, J. Saquer, and M. Ghosh, "Advancing mental disorder detection: A comparative evaluation of Transformer and LSTM architectures on social media," in Proc. 2025 IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC), Toronto, ON, Canada, 2025, pp. 193–202, doi: 10.1109/COMPSAC65507.2025.00033.

[14] N. Hussain, A. Qasim, G. Mehak, M. Zain, G. Sidorov, A. Gelbukh, and O. Kolesnikova, "Multi-level depression severity detection with deep Transformers and enhanced machine learning techniques," AI, vol. 6, no. 7, Art. no. 157, 2025, doi: 10.3390/ai6070157.

[15] N. Wawale, R. Owhal, S. Mane, and D. Shivankhede, "Hybrid mental health detection framework: Combining domain-specific BERT with psycholinguistic features for mental health detection," 2026.

[16] C. Xin and L. Q. Zakaria, "Integrating BERT with CNN and BiLSTM for explainable detection of depression in social media contents," IEEE Access, vol. 12, pp. 161203–161212, 2024, doi: 10.1109/ACCESS.2024.3488081.

[17] I. Mohammed and H. Hassani, "Mining mental health signals: A comparative study of four machine learning methods for depression detection from social media posts in Sorani Kurdish," arXiv preprint arXiv:2508.15829, 2025.

[18] J. Sun, S. Ma, Y. Fan, and P. Washington, "Evaluating large language models for anxiety and depression classification using counseling and psychotherapy transcripts," arXiv preprint arXiv:2407.13228, 2024.

[19] N. Raihan, S. S. C. Puspo, S. Farabi, A.-M. Bucur, T. Ranasinghe, and M. Zampieri, "MentalHelp: A multi-task dataset for mental health in social media," in Proc.

2024 Joint Int. Conf. Computational Linguistics, Language Resources and Evaluation (LREC-COLING), Torino, Italy, 2024, pp. 11196–11203.

[20] K. Hasan and J. Saquer, “A benchmark suite of Reddit-derived datasets for mental health detection,” arXiv preprint arXiv:2604.23458, 2026.