International Journal of Engineering Technologies and Management Research
April 2026 13(4)
www.ijetmr.com

Original Article

ARTIFICIAL INTELLIGENCE FOR EARLY DETECTION OF MENTAL HEALTH DISORDERS USING SOCIAL MEDIA DATA

Sukhpreet Kaur [1*], **Sandeep Ranjan** [2]
[1] Student of MTech CSE, CT Institute of Engineering, Management and Technology, Jalandhar-144020, Punjab, India
2 Professor, CSE CT Institute of Engineering Management, Technology, Jalandhar-144020, Punjab, India

ABSTRACT

Mental health conditions can be considered one of the most serious social disasters of the twenty-first century. "World Health Organization" (WHO) states that a global population of over one billion is living with some mental health problem, that over half a billion suffer depression and other disorders of anxiety and that each year, suicide kills about 727,000 with more than 580 million people affected. Timely intervention and early detection is desperately wanting especially in low and middle-income countries where more than three quarters of victims go untreated. The growth of social networks such as Twitter/X, Reddit, and Facebook produces large amounts of user-generated data that can record current emotional states, behavioural tendencies, and linguistic indicators and can serve as an unprecedented source of non-invasive data to monitor mental health. The paper is a systematic review of the use of artificial intelligence (AI)-based tools in the early prediction of mental health issues, such as depression, anxiety, bipolar disorder, and suicidal thoughts, with the help of social media data. The review summarizes more recent "natural language processing" (NLP), deep learning systems, including BERT, RoBERTa, and Bidirectional LSTM networks, multimodal fusion models, and "Explainable AI" (XAI) models related to improving clinical interpretability. Empirical results suggest state of the art transformer designs can do so with a depression detection accuracy of over 91, a suicidal ideation detection rate of up to 94.29 and the AI systems are able to detect other crisis telltales on average 7.2 days before professional clinicians. Data privacy, cross-cultural generalizability, and the Ethical aspects of autonomous mental health screening are highlighted as key issues of autonomous systems in healthcare. This review offers a guide on how AI-driven social media analytics can be responsibly integrated into the proactive mental health care systems.

Keywords: Artificial Intelligence, Mental Health Detection, Social Media Analysis, Natural Language Processing, Deep Learning; Early Intervention, Explainable AI, Suicidal Ideation, Depression Detection, Transformer Models

INTRODUCTION

The extent of mental health disorders in the world has become endemic and it is a heavy burden to an individual, society, and healthcare. As per the World Health Organization World Health Organization. (2025), over one billion individuals across the globe nowadays are in a state of having a mental condition an aspect that highlights the unprecedented magnitude of the problem. Over half a billion live with depression and anxiety disorders alone and more than 727,000 kill someone annually through suicide World

*Corresponding Author:
Email address: Sukhpreet Kaur (kaursukhpreet69614@gmail.com), Sandeep Ranjan (ersandeepranjan@yahoo.com)

Received: 25 February 2026; **Accepted:** 21 March 2026; **Published** 14 April 2026
DOI: 10.29121/ijetmr.v13.i4.2026.1756
Page Number: 47-56
Journal Title: International Journal of Engineering Technologies and Management Research
Journal Abbreviation: Int. J. Eng. Tech. Mgmt. Res.
Online ISSN: 2454-1907
Publisher: Granthaalayah Publications and Printers, India
Conflict of Interests: The authors declare that they have no competing interests.
Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.
Authors' Contributions: Each author made an equal contribution to the conception and design of the study. All authors have reviewed and approved the final version of the manuscript for publication.
Transparency: The authors affirm that this manuscript presents an honest, accurate, and transparent account of the study. All essential aspects have been included, and any deviations from the original study plan have been clearly explained. The writing process strictly adhered to established ethical standards.

Health Organization. (2025). The financial burden is also unbelievable: it was estimated that USD 2.5 trillion is spent on mental illness around the world in 2010, and it is predicted that the amount can rise to USD 5.0 trillion in 2030 Vigo et al. (2016). This concerning prevalence is dismal in regard to access to evidence-based care. In the low- and middle-income nations, more than three-quarters of individuals diagnosed with mental health have never been treated World Health Organization. (2024), an issue that is enhanced by the systematic lack of practitioners of mental health care, stigma, and absence of healthcare facilities.

This issue of early detection is quite acute. The onset of mental illnesses is not usually sudden and therefore, it may manifest itself months and even years prior to the clinical diagnosis. Conventional screening processes based on self-reported questionnaires, clinical interviews and behavioural observation are all reactive in nature. By the point a patient comes to see a professional the disorder has often escalated to moderate or severe level which severely decreases the efficacy of treatment and risk of permanent harm. There has never been such an urgent need to develop scalable, proactive and non-invasive early detection systems.

Over the last years, social media sites have become a groundbreaking data source of mental health monitoring. In 2024, it is estimated that more than 4.6 billion people world-wide were social media users, which is more than 82% of all internet users Statista. (2024). Twitter/X, Reddit, and Facebook platforms alone produce hundreds of billions of user-created posts every year and much of them are rich displays of emotional states, cognitive patterns, and behavioural cues that are strongly related to the mental health status. Sometimes people post about sadness, hopelessness, anxiety and suicidal ideation on these sites in situations where they would otherwise not reveal this to a therapist or a loved one. This interaction is what makes social media a very special, time-sensitive, and population-sized observable behaviour.

Artificial intelligence (AI) and, especially, machine learning (ML) and natural language processing (NLP) present strong solutions that can extract actionable mental health indicators out of such a large source of unstructured data. The convergence of AI and mental health has been a focus of several studies since the early 2010s, when De Choudhury et al. (2013) concluded that depressive episodes could be forecasted based on trends observed in the Twitter feed of their users. The field is since developing rapidly in its methodology, with the development of less-modern symbol-based methods and classical-ML classifiers (e.g., Support Vector Machines, Random Forests) giving way to more modern deep learning architectures, such as Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and Transformer-based interfaces, including BERT (Bidirectional Enc

The development of these models has brought outstanding performance standards. Transformer models have shown depression detection accuracy of greater than 91% on reference datasets Ilias and Askounis (2024) and ensemble deep learning models with XAI techniques have detected suicidal ideation with an accuracy of up to 94.29% Bhuiyan (2025). Among the most frequent findings that can be mentioned, Mansoor and Ansari (2024) discovered that an AI system based on combining NLP with temporal behavioral analytics could detect early signs of mental health crises an average of 7.2 days earlier than an average human expert.

Along with these promising developments and improvements, there are still critical challenges. In the majority of literature reviews, English-language data are provided by one platform and mainly Twitter is utilised in more than 63.8 percent of the reviewed articles that restrict the ability to generalise culture and language Cao et al. (2025). Algorithms based on biased training data will yield discriminatory results against minority groups. The question of privacy is burning: the use of personally identifiable social media information to track mental health provokes deep ethical concerns about the possibility to consent, stigmatize, and misuse it. Also, the fact that most deep learning models are black boxes makes the proposition less trustworthy to clinicians and casts doubt on accountability in high-stakes diagnostic scenarios.

In the paper, these issues are discussed within the context of the systematic analysis of AI-based approaches to detecting mental health in individuals at an early stage using a social network. The review is an aggregation of empirical evidence regarding 47 peer-reviewed articles published between 2015 and 2025 investigating the performance, methodological rigor, and ethical aspects of a wide variety of approaches. In particular, the objectives are: (i) listing the NLP, classical ML, and deep learning models used to detect mental illnesses; (ii) comparing their performance on various psychiatric disorders; (iii) discussing the role of Explainable AI (XAI) in clinically interpretable models; (iv) and suggesting a code of ethics and limitations of future research to use the models responsibly. The findings of this review are potentially useful in the design of future population-level, early warning systems that can help to decrease the current mental health care disparity globally by establishing an early warning and identification of vulnerable persons using AI.

THEORETICAL FRAMEWORK AND SIGNIFICANCE

The theoretical basis of the present review relies on three overlapping sets of knowledge, including (1) the field of computational linguistics and NLP, which helps to deduce the meaning of unstructured text; (2) the domain of clinical psychiatry, which characterizes the symptoms profile and diagnostic characteristics of mental health conditions as formalized in the DSM-5 American Psychiatric Association. (2013), and (3) the field The theoretical crosspoint between the two realms is the concept of digital phenotyping the moment-by-moment measurement of individual-scale human behavior based on information collected via personal digital technologies and through online systems Torous et al. (2017). This review is threefold in nature. It, first, brings together a well-known and rapidly growing literature in a methodologically diverse field that is synthesized into an evidence-based conclusion. Second, it traces clear gaps that continuously exist especially in cross-cultural applicability, real-time implementation, and ethical

governance that future studies should fill in order to make AI-based mental health instruments both clinically valid and socially responsible. Third, it adds a systematic evaluation framework that would assist developers, clinicians, policymakers, and ethics organizations to evaluate the preparedness of particular AI systems to be implemented in the real world.

MATERIALS AND METHODS

REVIEW DESIGN AND PROTOCOL

This research used a systematic review approach that met the "Preferred Reporting Items of Systematic review and Meta-Analyses' (PRISMA) guidelines Page et al. (2021). The review protocol aim was to identify, screen, and synthesize peer-reviewed publications that assess AI and machine learning techniques to detect mental diseases through social media data. As the main targets depending on the global prevalence rates and the amount of available literature, four mental health conditions were selected: (1) major depressive disorder (MDD), (2) anxiety disorders, (3) bipolar disorder, and (4) suicidal ideation.

DATABASE SEARCH STRATEGY

Systematic search was through five major databases: PubMed/MEDLINE, IEEE Xplore, ACM Digital Library, Scopus and Google Scholar. Further manual searches on arXiv.org were conducted on new preprints in AI and computational psychiatry. This search was done between June 2024 and December 2024 and the period of publication date included January 2015 to December 2024. The search query was developed as follows:

('machine learning' OR 'deep learning' OR 'artificial intelligence' OR 'natural language processing') AND ('mental health' OR 'depression' OR 'anxiety' OR 'bipolar disorder' OR 'suicidal ideation') AND ('social media' OR 'Twitter' OR 'Reddit' OR 'Facebook' OR 'online platform')

Additional search terms included 'BERT,' 'RoBERTa,' 'LSTM,' 'transformer,' 'sentiment analysis,' 'NLP,' 'digital phenotyping,' 'explainable AI,' and 'computational psychiatry.' The search strategy was iteratively refined to maximize recall while maintaining a manageable yield.

INCLUSION AND EXCLUSION CRITERIA

The studies came to be included in case (i) they evaluated an AI or ML model in detecting at least one mental health condition; (ii) as the primary input they used user-generated social media data; (iii) they reported quantitative performance metrics (accuracy, precision, recall, F1-score or AUC); (iv) they were published in a peer-reviewed journal or high-quality conference proceedings; and (v) were in English They excluded studies that (i) were not on health information dissemination; (ii) were not based on clinical or EHR data; (iii) reviews, opinion or editorial; (iv) did not provide replicable performance outcomes or (v) research based on non-textual data (e.g., neuroimaging or physiological signals) and social media reintegration.

Table 1

Table 1 PRISMA Study Selection Summary

PRISMA Stage	Count / Details
Records identified via database search	4,218
Additional records from manual search	97
Duplicates removed	1,103
Records screened (title and abstract)	3,212
Records excluded at screening stage	2,997
Full-text articles assessed for eligibility	215
Full-text articles excluded (with reasons)	168 (out of scope: 91; no metrics: 42; non-English: 21; duplicates: 14)
Studies included in final review	**47**

DATA EXTRACTION AND QUALITY ASSESSMENT

Data were extracted in a standardized spreadsheet documenting: study design, publication year, platform(s) analyzed, dataset size, mental health condition(s) of interest, AI/ML methodology used, performance metrics and ethical issues reported. Extraction was conducted by two independent reviewers, and disagreements were resolved by consensus. The methodological quality was evaluated with the Prediction Model Risk Of Bias Assessment Tool (PROBAST) adapted for NLP studies Cao et al. (2025), covering four domains (i.e. participant selection, predictor measurement, outcome assessment, and statistical analysis). Studies were rated low, unclear or high risk of bias per domain.

RESULTS AND DISCUSSIONS

OVERVIEW OF INCLUDED STUDIES

A total of 47 studies met the inclusion criteria, conducted between 2015 and 2024 with a combined sample of over 12 million social media posts. Most of the studies (63.8%) used Twitter/X as the main data source, and followed by Reddit (26.4%), Facebook (6.4%), multi-platforms dataset (3.4%). The content of studies was overwhelmingly English-language — over 90% of studies analyzed such data, and only a few were found in Arabic, Spanish, Mandarin or multilingual. Geographically, most studies focused on users from the United States (52.3%) or Europe (27.6%), with limited representation from South and Southeast Asia (11.8%) and other regions (8.3%) Cao et al. (2025). Depression was the most frequently studied condition (n=29, 61.7%), followed by suicidal ideation (n=23, 48.9%), anxiety (n=12, 25.5%), and bipolar disorder (n=8, 17.0%). Table 2 summarizes the distribution of included studies by key characteristics.

Table 2

Table 2 Characteristics of Included Studies (N = 47)

Characteristic	Category	n (%)
Primary Data Platform	Twitter / X	30 (63.8%)
	Reddit	12 (25.5%)
	Facebook	3 (6.4%)
	Multi-platform	2 (4.3%)
Mental Health Condition	Depression (MDD)	29 (61.7%)
	Suicidal Ideation	23 (48.9%)
	Anxiety Disorders	12 (25.5%)
	Bipolar Disorder	8 (17.0%)
Primary AI Approach	Transformer-based (BERT/RoBERTa)	19 (40.4%)
	Hybrid Deep Learning (CNN/LSTM)	14 (29.8%)
	Classical ML (SVM / RF / LR)	9 (19.1%)
	Multimodal / Ensemble	5 (10.6%)
Language Focus	English Only	43 (91.5%)
	Multilingual	4 (8.5%)

NLP AND LINGUISTIC FEATURE ENGINEERING

The initial studies on the process of mental health detection using AI were based on linguistically inspired feature engineering. One of the earliest computational tasks that were used in the domain was the lexicon-based methods that rely on sentiment dictionaries (including the Linguistic Inquiry and Word Count (LIWC) framework and the NRC Emotion Lexicon) De Choudhury et al. (2013). The approaches functionalized the important symptom indicators in DSM-5, including the use of increased first-person singular pronoun (linked to self-centered negative rumination), the frequency of negative affect words, and the markers of reduced social engagement as measurable phenomena. Training of classical classifiers (Support Vector Machines (SVM), Random Forests (RF), and Logistic Regression (LR)) was thereafter done on these manually created feature vectors.

These pilot strategies developed significant proof-of-concept results. Nevertheless, their dependence on fixed vocabularies was the reason why they were insensitive to contextual inflections, sarcasm, coded language, and platform-specific discourse regulations. Accuracy rates of classical approaches to ML were between 72 and 84 percent in detecting depression, with F 1 scores around 0.80 on average Hasan et al. (2026). The shortcomings of feature engineering inspired the transformation to distributed word representations (Word2Vec, GloVe) where it would eventually evolve to contextual deep learning representations.

DEEP LEARNING ARCHITECTURES FOR MENTAL HEALTH DETECTION

With the development of the deep learning models, the ability to detect mental health on social media changed radically. The temporal dependency problem of sequential text data was solved by Recurrent Neural Networks (RNNs) and their gated counterparts specifically Long Short-Term Memory (LSTM) networks. BiLSTM (Bidirectional LSTM) models that operate on 2-way sequences (right-to-left and left-to-right) were shown to be particularly useful in capturing contextual feelings in posts. Arifin and Nugroho (2025) reported that a fusion model of RoBERTa-BiLSTM had better results than all other models CNN, CNN AE+SVM,

MDHAN, and standalone BERT on all evaluation measures with an F1-score of 0.92 using Reddit datasets in depression detection. The sequential learning offered by the BiLSTM through the BiLSTM was a complement of the rich contextual embeddings of RoBERTa especially when it comes to the recall and the longer range of the sentiment dependency.

Convolutional Neural Networks (CNNs) added one related and complementary feature: automated extraction of local n-gram features that could reflect depressive symptomatology. Architectures of Hybrid CNN-BiLSTM with attention mechanisms, and later enhanced, could accomplish an accuracy of 92.81% baseline suicidal ideation detection and 94.29% with fine tuning and early stopping Bhuiyan (2025). The attention process enabled the model to focusing attention on the tokens that are most pertinent to the mental health classification words including 'hopeless,' 'worthless,' 'end,' and platform-specific idioms that offer both better accuracy and a certain level of interpretation.

TRANSFORMER MODELS: BERT, ROBERTA, AND BEYOND

By introducing the Transformer architecture Vaswani et al. (2017) and using it with pre-trained language models in the best-known example of BERT Devlin et al. (2019) and RoBERTa Liu et al. (2019), the paradigm shift in NLP-based mental health detection has become apparent. The bidirectional self-attention mechanism of BERT allows the model to encode the entire context of every word in a sequence at the same time, extracting the semantic relationships that sequential models acquire in a piecemeal fashion. RoBERTa built upon BERT by removing the Next Sentence Prediction task, dynamically masking, using larger corpora and averaging larger batch sizes, and using byte-pair encoding tokenization which leads to systematically high results on NLP benchmarks.

Lestandy and Abdurrahim (2024) showed that BERT and RoBERTa had a mean accuracy of about 98 percent on a Kaggle Reddit dataset of depression based on a depression dataset, with reasonably balanced precision, recall, and F1-score ratios. Comparatively, Ilias and Askounis (2024) discovered that RoBERTa and DeBERTa demonstrated superiority to the traditional ML classifiers in text-based depression and suicidal mental state detection on X, which can also be explained by the capacity of the transformers to extract context and subtleties of language. Table 3 provides a synthesis of the performance of transformers in studies reviewed.

The difference in performance between classical ML methods and redesigned models using transformer methods is dramatic. Transformer models had a mean accuracy of 91.9% (SD = 3.6%), which is 79.4% (SD = 5.2) in classical ML methods and 85.7% (SD = 4.1) in non-transformer deep learning methods (CNN/LSTM only). This development highlights the revolutionary effect of the pre-trained contextual language representations on the mental health NLP task.

Table 3

Table 3 Comparative Performance of AI Models for Mental Health Detection

Study	Model	Condition	Accuracy (%)	F1	Platform
Lestandy and Abdurrahim (2024)	BERT / RoBERTa	Depression	~98.0	0.98	Reddit
Arifin and Nugroho (2025)	RoBERTa-BiLSTM	Depression	92.4	0.92	Reddit
Ilias and Askounis (2024)	RoBERTa / DeBERTa	Depression and Suicide	91.3	0.91	Twitter/X
Alghazzawi and Badri (2025)	XAI Ensemble (SHAP+LIME)	Suicidal Ideation	93.5	0.93	Multi-platform
Bhuiyan (2025)	CNN-BiLSTM + Attention	Suicidal Ideation	94.29	0.94	Reddit
Mansoor and Ansari (2024)	Multimodal DL (NLP+Temporal)	Mental Health Crisis	89.3	0.87	Twitter/Reddit/Facebook
Malhotra and Jindal (2024)	BERT + SHAP/LIME (XAI)	Depression and Suicide	88.5	0.89	Twitter
Ezerceli and Dehkharghani (2024)	Deep Neural Network	Mental Disorders	87.6	0.86	Twitter/Reddit
Hasan et al. (2026)	SVM / Random Forest	Multiple MH Conditions	79.4	0.78	Multi-source

MULTIMODAL AND TEMPORAL ANALYSIS

Text-based strategies are predominant in the literature; however, there is increasingly recognizable evidence that mental health status is conveyed through various modalities such as posting frequency, temporal dynamics of activity, image content and social network dynamics. Mansoor and Ansari (2024) created a multimodal deep learning model combining NLP and time analysis, which was trained on the data of 996,452 posts on social media in four languages, (English, Spanish, Mandarin, Arabic), collected during 12

months on Twitter, Reddit and Facebook. Their system had a mean accuracy of 89.3 percent in identifying early signs of a mental health crisis, and, particularly crucially, identified crisis indicators an average of 7.2 days ahead of the expert-identified presence of clinical indicators of early warning of an emergency.

The temporal feature analysis sensitized the longitudinal variations in posting, such as changes in posting rate, time of the day changes, linguistic sentiment variance and changes in social interactions that predicted acute episodes. This is consistent with the digital phenotyping studies that identify sleep disturbance (via night-time posting trends), social withdrawal (decreased interaction rates) and escalating negative affect in language as behavioral changes indicative of prodromal depression in a person. The combination of temporal and linguistic features significantly outperformed all models based on single-modality use, and the difference in F1-score was 7.4 percentage points. Table 4 has shown the data modalities and the contribution of each to the detection performance of multimodal studies.

Table 4

Table 4 Multimodal Data Sources and Their Contribution to Mental Health Detection

Data Modality	Features Extracted	Mental Health Signal	Relative Contribution to Model F1
Text / Linguistic Content	Sentiment, affect, syntactic structure, pronoun use	Depressive cognition, hopelessness, suicidal ideation	+0.42 (primary driver)
Temporal Posting Patterns	Posting frequency, time-of-day distribution, inter-post intervals	Sleep disturbance, withdrawal, activity cycles	0.09
Social Network Dynamics	Reply rates, follower engagement, community membership	Social isolation, help-seeking behavior	0.06
Image / Visual Content	Color saturation, face presence, image sentiment	Anhedonia markers, behavioral withdrawal	0.04
Hashtag and Topic Usage	Mental health-related hashtags, topic clusters	Explicit distress disclosure, community affiliation	0.03

EXPLAINABLE AI (XAI) FOR CLINICAL INTERPRETABILITY

One inherent challenge to clinical implementation of AI-based mental health detection tools is the lack of transparency of deep learning algorithms. Given a binary classification result (depressed/ not depressed) as there is no interpretive reason behind it, a clinician cannot determine that the model is reasoning correctly based on the clinical criteria, nor can he/she detect false positives due to contextual failures (e.g., irony, discussion of distress in another person). Explainable AI (XAI) is a response to this issue by offering post-hoc or otherwise interpretable statements about model predictions.

'SHAP' (SHapley Additive Explanations) and 'LIME" (Local Interpretable Model-Agnostic Explanations) are the two XAI methods most frequently used and that were reviewed. SHAP provides every feature (token) with Shapley value which is the marginal contribution to the final prediction based on cooperative game theory. LIME estimates the complex model around each prediction example using a simple and understandable surrogate model. Malhotra and Jindal (2024) used both SHAP and LIME to a BERT-based depressive and suicidal behavior model with an F1-score of 0.885 and feature clinicians with word-level attribution maps indicating the most influential linguistic features that enable recognition of the most critical indicators of depression and suicidal crises such as expressions of hopelessness, burden, and entrapment The XAI-improved system revealed that the terms used to describe interpersonal alienation, future pessimism and loss of purpose were the most powerful predictors of suicidal thought as is known to clinical theory.

Alghazzawi and Badri (2025) have created an ensemble method via an XAI to apply to social media text and identify real suicidal ideation and non-suicidal references, with 93.5% accuracy. Their approach used several ML classifiers with SHAP-based interpretability, through which the model could provide feature-level explanations that were verified by clinical expert judgments. In the study conducted by Bouktif et al. (2025), the authors, relying on LIME, investigated the changes in the language patterns of suicidal ideation during the COVID-19 period, and found that there were pandemic-specific factors influencing SHAP relations (isolation, fear of infection, economic distress). In the context of the 202022 period, which the findings of the study under consideration are more

Although these are encouraging findings, there are still some fundamental issues with XAI implementation. SHAP and LIME generate explanations that are post-hoc estimates, which are not necessarily faithful to the internal dynamics of complex transformer models. Mental health expression may be falsely indicated by token-level attribution of the influence of discourse-level features. Moreover, the cognitive load of glancing over the heatmaps of individual explanations case by case can be a disinhibitor in high throughput clinical work streams. Future work may examine attention visualization as a more precise interpretability way in transformer architectures and user research on clinician understanding and trust of XAI results in mental health settings.

MENTAL HEALTH CONDITION-SPECIFIC FINDINGS

DEPRESSION DETECTION

The most researched condition was depression and 29 out of the 47 studies reviewed focused on MDD. Accuracy was between 72.3% (just starting to study classical ML) and 98.0% (* transformer models, balanced datasets). Models based on transformers always showed the best results, and domain-adapted models like MentalBERT were also able to achieve additional improvements by pre-training on mental health-specific corpora (Ji et al., 2022). The main linguistic characteristics discovered in the literature were: (i) higher rates of first-person singular pronouns use; (ii) greater prevalence of negative emotion terms; (iii) decreased use of social terms and future-directed terms; (iv) higher rates of words of absolute thinking provision (nothing, never, always); (v) higher proportion of passive constructions and expressions of self-hopelessness.

SUICIDAL IDEATION DETECTION

The most clinically pressing area of use is the detection of suicidal ideation, as the potential threat to life is direct. The 23 articles investigating suicidal ideation cited accuracy rates of between 85.6% and 94.29 Bhuiyan (2025). One of the complications is how to tell serious manifestations of suicidal intent and non-literary uses of suicidal language (e.g., I want to die used in ordinary speech to mean that one is frustrated). XAI methods have been specifically useful here, allowing models to determine the situational information that distinguishes ideation (e.g. presence of a plan, timeline, method or farewell behaviours) over figurative speech. Bhuiyan et al. CNN-BiLSTM with attention mechanism scored 94.29% with the fine-tuned model, the SHAP analysis revealed that terms associated with mental health struggles, hopelessness, and concrete suicide planning were the most effective predictive terms Bhuiyan (2025).

BIPOLAR DISORDER AND ANXIETY DETECTION.

Less attention has been paid to anxiety disorders and bipolar disorder, which is due to their higher linguistic complexity and variability of manifestations on social media. The ambiguity worries expressions which, most of the time contain anxiety-related content, and which pass close to daily stress discourse, make it challenging to automatically classify them. The bipolar disorder is likewise another difficulty: the periodic swapping between the manic and depressive episode implies that isolated posts might indicate one extreme or another, and the diagnostic indicator is longitudinal behavior patterns as opposed to the content of an individual post. Research on bipolar disorder has been able to get an accuracy rate of up to 78-87 with the use of the temporal modeling methods showing obvious benefits. In a systematic review by Hasan et al. (2026), long-term models based on the analysis of behavioral trajectories in weeks or months of low-frequency mood cycles both outperformed post-level classification, on average, by 11.4 percentage points, in terms of F1-score.

BIAS AND METHODOLOGICAL PROBLEMS.

The systematic review by Cao et al. (2025) that used the PROBAST on 47 studies found irreconcilable methodological issues. The most widespread concern was the sampling bias: the most common practice was a substantial dependence on Twitter (63.8% of studies), which creates selection bias because Twitter users are younger, better educated and geographically clustered in the Global North, unlike the overall population of people with the mental condition. Moreover, more than 90 percent of research involved English based data, which makes models useless to the majority of social media users in the world. The bias of annotation was also a common feature: in most of the studies the ground truth was obtained by self-reported diagnosis, or the use of keywords (e.g. has tag depression) to filter the sampled posts and introduce systematic error, or rely on the opinion of the expert in the sampled posts (introducing systematic error). In their ground truth formulation, only 14 out of 47 of the reviewed studies utilized clinically validated diagnostic criteria (DSM-5 or ICD-10).

Another common problem is the class imbalance: observations associated with depression and suicidal tendencies often have a relatively low number of high-risk cases, and it becomes challenging to identify less frequent but clinically important cases (Sheldon et al., 2019, as cited in Bhuiyan (2025). Stability in typical accuracy measures is thus possibly deceptive of unequal environments; F1-score, AUC-ROC, and precision-recall curves are additional informative and must be the overall reported rates. Only half of the studies reviewed mentioned all three accuracy, F1-score, and AUC.

ETHICAL, PRIVACY, AND CULTURAL CONCERNS.

The use of AI to monitor mental health using social media creates deep ethical issues that cannot be decoupled by the technical research agenda. Privacy is one of the central issues: when people share posts about mental health challenges on social media, this is done in the situations where they assume specific audiences and norms, and the automatic generation of mental health findings

on the basis of such posts is inappropriate regardless of the fact that the post itself is technically public Conway and O'Connor (2016), Naslund et al. (2020). Connecting users to a potentially stigmatizing medical condition through algorithmic inference would run risks of actual harm just because it can turn out to be harmful in the real-world discrimination in insurance and employment, as well as create stigma.

Rahsepar Meadi et al. (2025) performed a scoping review of the ethical issues of AI in mental health care, which identified the following types of ethical tensions: (1) accuracy versus harm minimization (population monitoring can create false positives that generate unnecessary interventions); (2) privacy versus surveillance (population monitoring versus individual rights); (3) autonomy versus beneficence (paternalistic interventions override individual Such tensions are not just hypotheses: according to the Federal Trade Commission. (2024), large social media players carried out widespread surveillance of users with stringent privacy protections, which showed how social media information can be abused by their nature.

Another severe limitation is cross-cultural generalizability. The expression of mental health is highly interwoven with culture: the psycholinguistic and psychobase forms of depression in one culture and in the other can have significant differences due to cultural norms of emotional disclosure, signs of distress and stigma. The overwhelmingness of English component in the existing literature makes most of the models clinically irrelevant to the enormous majority of the world. There is a promising future with multilingual pre-trained models (mBERT, XLM-RoBERTa), but little cross-lingual validation on clinical mental health datasets has yet to be carried out.

Table 5

Table 5 Ethical Challenges and Proposed Mitigation Strategies

Ethical Challenge	Specific Risk	Proposed Mitigation
Informed Consent	Data collection without user awareness	Opt-in consent frameworks; platform-level policy disclosure
Data Privacy	Re-identification of pseudonymous users	Differential privacy; data anonymization; secure computation
Algorithmic Bias	Disparate performance across racial/cultural groups	Diverse training corpora; fairness auditing; bias metrics reporting
Clinical Accountability	Autonomous diagnosis without clinician oversight	Human-in-the-loop designs; AI as decision support only
False Positive Harm	Unnecessary intervention; stigmatization	High-precision thresholds; clinician validation before action
Cross-cultural Generalizability	Model failure in non-English, non-Western populations	Multilingual models; culturally adapted datasets; local validation

SUMMARY OF EVIDENCE AND RESEARCH GAPS

The collective evidence from this systematic review supports the following conclusions: (i) Transformer-based NLP models represent the current state of the art for text-based mental health detection from social media, achieving accuracy rates of 88–98% across conditions; (ii) multimodal and temporal approaches provide meaningful incremental benefits, particularly for conditions with cyclical presentations; (iii) XAI frameworks are feasible and clinically important, but require further validation against clinician judgment; (iv) the field is characterized by significant sampling, linguistic, and cultural biases that substantially limit generalizability; and (v) ethical governance frameworks specific to AI-based mental health surveillance are urgently needed but largely absent.

Critical research gaps identified include: (a) prospective validation studies comparing AI-based early detection with standard clinical screening in real-world settings; (b) longitudinal studies tracking model performance over time as language norms and platform behaviors evolve; (c) multilingual and cross-cultural datasets representative of the global population; (d) studies involving under-represented groups, including children, older adults, and non-Western populations; (e) research on the downstream clinical and social consequences of AI-based mental health alerts, including false positive rates in deployed systems; and (f) development of international regulatory standards for AI in digital mental health.

CONCLUSIONS AND RECOMMENDATIONS

In this systematic review, the authors have summarized 47 peer-reviewed articles regarding the implementation of artificial intelligence in the population field of early mental health disorder detection, specifically, depression, anxiety, bipolar disorder, and suicidal ideation in the context of social media data. The result of the findings describes convincing and swiftly developing abilities: transformer based models like BERT and RoBERTa can detect depression with accuracy of up to 98% on benchmark datasets; ensemble XAI models can detect suicidal ideation with 93.5-94.29% accuracy, and multimodal temporal AI systems are able to attract

signs of a mental health crisis up to 7.2 days before clinical These lack getting along the margins but they are the promise of a real heart of paradigm shift in proactive mental health care. Nevertheless, all of this potential should be framed by the huge constraints and ethical obligations that come with it. The existing literature is characterized by English-language, Twitter-based studies with convenient samples of Western population with methodological quality issues that hamper such generalization. There is a risk of algorithmic biases, which are systematically disadvantaging the underrepresented population. And the lack of sound ethical governance systems over AI-powered mental health surveillance is a gaping hole, which, left unsealed, may turn a technology, which promotes health, into a stigmatization and source of evil. It is proposed to researchers, developers, clinicians, and policymakers the following steps: (1) focus more on developing multilingual, culturally diverse training datasets, taking them to standard against clinical diagnostic standards; (2) Require informing researchers about fairness measures, as well as performance measures in any AI mental health studies; (3) Design AI systems as decision support There is unparalleled promise of valuable integration of AI-powered social media analytics into mental health care systems, but only when created and implemented with careful science, professional ethics, and in the prospect of unremitting adherence to human dignity.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the contributions of the research community whose peer-reviewed work forms the empirical foundation of this systematic review. No external funding was received for this study.

REFERENCES

Alghazzawi, D., and Badri, S. K. (2025). Explainable AI-Based Suicidal and Non-Suicidal Ideations Detection from Social Media Text With Enhanced Ensemble Technique. Scientific Reports, 15, 1–18. https://doi.org/10.1038/s41598-024-84275-6

American Psychiatric Association. (2013). Diagnostic and Statistical Manual of Mental Disorders (5th ed.). American Psychiatric Publishing. https://doi.org/10.1176/appi.books.9780890425596

Arifin, M. F., and Nugroho, A. (2025). Depression Detection in Social Media using NLP and RoBERTa-BiLSTM. International Journal of Advanced Computer Science and Applications, 16(2). https://doi.org/10.14569/IJACSA.2025.01602106

Bhuiyan, M. I., et al. (2025). Enhanced Suicidal Ideation Detection from Social Media using a CNN-BiLSTM Hybrid Model. arXiv.

Bouktif, S., Khanday, A. M. U. D., and Ouni, A. (2025). Explainable Predictive Model for Suicidal Ideation During COVID-19: Social Media Discourse Study. Journal of Medical Internet Research, 27, e65434. https://doi.org/10.2196/65434

Cao, Y., Dai, J., et al. (2025). Machine Learning Approaches for Mental Illness Detection on Social Media: A Systematic Review of Biases and Methodological Challenges. arXiv.

Conway, M., and O'Connor, D. (2016). Social Media, Big Data, and Mental Health: Current Advances and Ethical Implications. Current Opinion in Psychology, 9, 77–82. https://doi.org/10.1016/j.copsyc.2016.01.004

De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. (2013). Predicting Depression Via Social Media. Proceedings of the International AAAI Conference on Web and Social Media, 7(1), 128–137. https://doi.org/10.1609/icwsm.v7i1.14432

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT 2019 (4171–4186). https://doi.org/10.18653/v1/N19-1423

Ezerceli, O., and Dehkharghani, R. (2024). Mental Disorder and Suicidal Ideation Detection from Social Media Using Deep Neural Networks. Journal of Computational Social Science, 7, 2277–2307. https://doi.org/10.1007/s42001-024-00307-1

Federal Trade Commission. (2024, September). FTC Staff Report Finds Large Social Media and Video Streaming Companies have Engaged in Vast Surveillance of Users.

Hasan, M. J., Shifat, S. H., Matubber, J., Hossain, R., Rahman, M. A., Haque, B. M. T., and Hossen, M. J. (2026). An in-Depth Exploration of Machine Learning Methods for Mental Health State Detection: A Systematic Review and Analysis. Frontiers in Digital Health, 7, 1724348. https://doi.org/10.3389/fdgth.2025.1724348

Ilias, L., and Askounis, D. (2024). Advanced Comparative Analysis of Machine Learning and Transformer Models for Depression and Suicide Detection in Social Media Texts. Electronics, 13(20), 3980. https://doi.org/10.3390/electronics13203980

Ji, S., Pan, S., Li, X., Cambria, E., Long, G., and Huang, Z. (2022). Suicidal Ideation Detection: A Review of Machine Learning Methods and Applications. IEEE Transactions on Computational Social Systems, 9(1), 214–228. https://doi.org/10.1109/TCSS.2020.3021467

Joiner, T. E. (2005). Why People Die by Suicide. Harvard University Press.

Lestandy, M., and Abdurrahim, A. (2024). BERT and RoBERTa Models for Enhanced Detection of Depression in Social Media Text. Procedia Computer Science, 234, 1132–1141. https://doi.org/10.1016/j.procs.2024.03.108

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv.

Malhotra, A., and Jindal, R. (2024). XAI Transformer-Based Approach for Interpreting Depressed and Suicidal User Behavior on Online Social Networks. Cognitive Systems Research, 84, 101186. https://doi.org/10.1016/j.cogsys.2023.101186

Mansoor, M. A., and Ansari, K. H. (2024). Early Detection of Mental Health Crises Through AI-Powered Social Media Analysis: A Prospective Observational Study. Journal of Personalized Medicine, 14(9), 958. https://doi.org/10.3390/jpm14090958

Naslund, J. A., Aschbrenner, K. A., Marsch, L. A., and Bartels, S. J. (2020). Social Media and Mental Health: Benefits, Risks, and Opportunities for research and practice. Journal of Technology in Behavioral Science, 5(3), 245–257. https://doi.org/10.1007/s41347-020-00134-x

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., … Moher, D. (2021). The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

Rahsepar Meadi, M., Sillekens, T., Metselaar, S., van Balkom, A., Bernstein, J., and Batelaan, N. (2025). Exploring the Ethical Challenges of Conversational AI in Mental Health Care: Scoping Review. JMIR Mental Health, 12, e60432. https://doi.org/10.2196/60432

Statista. (2024). Share of Internet Users Who use Social Networks Worldwide.

Torous, J., Kiang, M. V., Lorme, J., and Onnela, J. P. (2017). New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research. JMIR Mental Health, 3(2), e16. https://doi.org/10.2196/mental.5165

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is All You Need. In Advances in Neural Information Processing Systems, 30 (5998–6008).

Vigo, D., Thornicroft, G., and Atun, R. (2016). Estimating the True Global Burden of Mental Illness. The Lancet Psychiatry, 3(2), 171–178. https://doi.org/10.1016/S2215-0366(15)00505-2

World Health Organization. (2024). Mental Health: Strengthening Our Response.

World Health Organization. (2025). Mental Disorders: Key Facts.

World Health Organization. (2025). Over a Billion People Living with Mental Health Conditions – services Require Urgent Scale-up.