JIRBDAI

Journal of Institutional Research, Big Data Analytics and Innovation

A Comprehensive Review of Machine Learning Algorithms for Depression Prediction with a Fuzzy Logic Perspective

Oladejo Qazeem Ayokunle[1], Aro Taye Oladele[2], Ejidokun Adekunle Olugbenga[3], Owonipa Ayodeji Abimbola[4], Buraimoh Olawale Henry[5], Lawal Hussein Taiye[6]

1,2,5, Department of Computer Science, Al-Hikmah University, Ilorin, Nigeria
[3]Kampala International University, Kampala, Uganda
[4]School of Informatics, University of Edinburgh, United Kingdom.
[6]National Open University of Nigeria, Ilorin, Kwara State, Nigeria

connectwithqazeem@gmail.com[1], taiwo774@gmail.com, gbengaejidokun@gmail.com, owonipaa@gmail.com, Hlawal@noun.edu.ng

ABSTRACT Depression is a pervasive mental-health disorder characterized by persistent low mood, anhedonia, cognitive impairment, and somatic symptoms that impair daily functioning. In this systematic review we examined 162 peer-reviewed studies published between 2010 and 2024, identified via a PRISMA-guided search of PubMed, IEEE Xplore, Scopus, Web of Science and Google Scholar using keywords related to "depression", "prediction", "machine learning", "neuro-fuzzy" and "fuzzy logic"; studies were included if they reported empirical predictive models on human subjects, supplied quantitative performance metrics, and were written in English, and were screened independently by two reviewers with data extraction covering algorithms, modalities, evaluation metrics, interpretability and ethical considerations. We synthesize advances across EHR analyses, NLP of clinical notes and social media, multimodal audio–visual signals, and physiological monitoring, and identify persistent gaps, most notably a shortage of transparent, generalizable models and a near-absence of systematic evaluations of fuzzy or hybrid neuro-fuzzy approaches in the depression domain. Unlike previous reviews that focus narrowly on performance or a single modality, this paper is the first to quantify and compare interpretability, generalizability and ethical risk across traditional ML and fuzzy/hybrid models, highlight where fuzzy inference can encode clinical expertise to manage symptom uncertainty, and propose a practical, hybrid neuro-fuzzy framework and evaluation checklist aimed at producing accurate, explainable and ethically responsible depression screening tools deployable in real-world clinical settings.

KEYWORDS: Depression, Prediction, Fuzzy Logic, Machine Learning, Mental Health, and Natural Language Processing (NLP)

I. INTRODUCTION

Depression is a common, disabling mental-health disorder characterized by persistent low mood, loss of interest or pleasure, cognitive disturbances, and somatic symptoms that markedly reduce an individual's capacity to function in work, study, and relationships (World Health Organization, 2023; National Institute of Mental Health, 2024). It is a leading contributor to the global burden of disease, associated with increased risk of chronic physical illness, suicidal behavior, and substantial productivity losses; these population-level harms underline an urgent need for more effective early-detection and screening strategies that can operate at scale and in diverse clinical and community settings (Pan American Health Organization, 2021; Rong et al., 2025). Over the past decade, machine learning (ML) has emerged as a powerful tool for depression prediction, leveraging electronic health records, natural-language signals from clinical notes and social media, multimodal audio–visual cues, and physiological sensors to detect patterns that traditional

methods miss. However, high reported accuracy often coexists with limited interpretability, reduced generalizability across populations and settings, and unresolved ethical concerns around privacy, bias, and clinical accountability, barriers that have so far constrained translation into routine care. Fuzzy logic and hybrid neuro-fuzzy approaches offer a complementary pathway by encoding clinical expertise as graded, human-readable rules and explicitly modelling uncertainty; they have the potential to make ML-based predictions more transparent, clinically meaningful, and robust to noisy or heterogeneous data. Despite these advantages, systematic evaluations of fuzzy and hybrid methods within the depression-prediction literature remain sparse, and existing reviews typically emphasize performance metrics or single modalities rather than interpretability, deployment readiness, and ethical risk. This review, therefore, aims to bridge that gap by synthesizing evidence across modalities, comparing traditional ML and fuzzy/hybrid approaches on interpretability and generalizability, and proposing practical directions for research and clinical deployment.

To guide the review, we address five research questions/objectives:

RQ1: What machine learning approaches and data modalities have been used for depression prediction, and with what empirical performance?

RQ2: What are the major challenges in interpretability, generalizability, and ethical risk within existing depression-prediction studies?

RQ3: How can fuzzy logic and hybrid neuro-fuzzy models be integrated with ML to improve transparency, handle symptom uncertainty, and encode clinical expertise?

(a). What evaluation practices, datasets, and metrics are necessary to assess both predictive accuracy and real-world readiness (interpretability, fairness, privacy, portability)?

(b). Based on current evidence, what practical hybrid architectures, validation strategies, and research priorities will most effectively move interpretable, ethical depression-prediction systems toward clinical adoption?

II. LITERATURE REVIEW

A. Scope and Methodology

This review adopted a systematic, reproducible search strategy to identify empirical studies on machine-learning (ML) approaches for depression prediction published between 2010 and 2024. We searched PubMed, IEEE Xplore, Scopus, Web of Science, ACM Digital Library and Google Scholar (initial search date: March 2025) using combinations of keywords and Boolean operators such as “depression”, “major depressive disorder”, “prediction”, “machine learning”, “deep learning”, “natural language processing”, “neuro-fuzzy”, and “fuzzy logic”. Inclusion criteria were: (1) peer-reviewed empirical studies that developed or evaluated predictive models for depression on human subjects; (2) provision of quantitative performance metrics (e.g., AUC, accuracy, F1); (3) English language; and (4) sufficient methodological detail to extract data. Exclusion criteria included narrative reviews, opinion pieces, studies without primary data or predictive models, and animal studies. Titles and abstracts were screened independently by two reviewers; full texts were assessed for eligibility, and data were extracted into a pre-specified form covering study aims, population, data modality, sample size, algorithms, validation strategy, performance metrics, interpretability methods, and reported limitations. The final set comprises 162 studies, and results are synthesized thematically.

B. Type of Depression

Major Depressive Disorder (MDD) is the prototypical form of depressive illness, characterized by a constellation of symptoms including a persistently low mood, marked anhedonia, loss of interest or pleasure in most activities, significant changes in weight or appetite, sleep disturbances, psychomotor agitation or retardation, fatigue, diminished ability to think or concentrate, and recurrent thoughts of death or suicide.

Volume 1, Issue 3, 2025
https://doi.org/10.5281/zenodo.16995362

MDD affects over 300 million people worldwide and has been projected by the World Health Organization to become the leading cause of global disease burden by 2030 (Cui et al., 2024). Diagnosis requires the presence of at least five of these symptoms for a minimum of two weeks, with at least one symptom being depressed mood or anhedonia, and exclusion of manic or hypomanic episodes per DSM-5 criteria (StatPearls, 2024) (NCBI).

Persistent Depressive Disorder (PDD), previously known as dysthymia, is a chronic form of depression characterized by a depressed mood for most of the day, for more days than not, over at least two years in adults or one year in children and adolescents. Introduced in DSM-5 (2013) to unify dysthymic disorder and chronic major depressive disorder, PDD features less severe but more enduring symptoms than MDD, often leading to substantial functional impairment over time (StatPearls, 2023) (NCBI). Individuals with PDD may experience intermittent major depressive episodes ("double depression"), further complicating the clinical course and treatment response.

Seasonal Affective Disorder (SAD) is a subtype of depression exhibiting a seasonal pattern, most commonly with onset in late autumn or early winter and remission in spring. SAD manifests with many core depressive symptoms, low mood, hypersomnia, increased appetite, often craving carbohydrates, weight gain, and social withdrawal, but its temporal pattern distinguishes it from non-seasonal depressions. (StatPearls, 2023) reports that SAD prevalence ranges up to 10% in high-latitude populations, underscoring the role of seasonal light variation and circadian rhythm dysregulation in its pathophysiology (NCBI).

Postpartum Depression (PPD), also termed peripartum depression when including pregnancy onset, refers to a major depressive episode with onset during pregnancy or within four weeks after childbirth, though risk persists for up to one year postpartum. Global PPD prevalence is around 17.7%, varying by socioeconomic context, and emphasizes its serious consequences for both maternal and infant health, including impaired bonding, adverse developmental outcomes, and elevated risk of chronic mood disorders (Dimcea et al., 2024). PPD's definition aligns with DSM-5 criteria for MDD with peripartum onset and necessitates targeted screening and early intervention to mitigate long-term impacts.

C. Depression Detection

Depression detection refers to the automated identification of depressive states in individuals through computational analysis of diverse data sources. By leveraging machine-learning algorithms, these systems aim to recognize patterns indicative of depression, such as linguistic markers, physiological signals, or behavioral cues, without relying solely on clinician judgment. Automated depression assessment tools assist in objective diagnosis and remote care to improve the quality of mental healthcare, moving beyond traditional semi-structured interviews that are often subjective and resource-intensive (Mao, Wu, and Chen, 2023). Similarly, machine learning models applied to electronic health records can achieve classification performances with AUC values averaging around 0.8 for depression prediction, demonstrating the feasibility of data-driven detection in clinical settings (Nickson et al., 2023). The impetus for developing automated detection systems stems from the high global burden of depression and the barriers to timely diagnosis. Depression affects hundreds of millions worldwide, yet over 75% of affected individuals in low- and middle-income countries lack access to adequate care (Mao et al., 2023). Automated tools provide a scalable solution that can flag at-risk individuals early, potentially before they engage with mental health services, thereby enabling prompt intervention. In a systematic review of EEG-based diagnostic approaches, there is an emphasis that machine intelligence can transform subjective assessments into objective biomarkers, underscoring the potential for continuous, passive monitoring in both clinical and everyday environments (Nassibi et al., 2025).

Current depression detection methodologies span a range of modalities, including electronic health records (EHRs), natural language processing of social media or clinical notes, speech and facial expression analysis, and physiological signals such as EEG. Regression methods, support vector machines, and ensemble classifiers applied to EHR data achieved comparable performance to primary care assessments, while multimodal approaches, combining acoustic, semantic, and visual features, have shown promise in automated diagnosis frameworks (Nickson et al., 2023). However, as pointed out in the study of social

https://doi.org/10.5281/zenodo.16995362

media-based detection, biases in data sampling and methodological inconsistencies remain major hurdles, affecting model reliability and generalizability across populations (Cao et al., 2024). Despite these advances, several challenges persist in depression detection research. Model interpretability is critical for clinical adoption, yet many deep learning approaches operate as black boxes, limiting trust among healthcare practitioners. Generalizability remains an issue, as algorithms trained on specific datasets often underperform when applied to new contexts or demographic groups. Additionally, ethical considerations around privacy, consent, and data bias must be addressed to ensure equitable and responsible deployment. Overcoming these challenges will be essential to realizing the full potential of machine learning–based depression detection and integrating such tools into routine mental health care.

III. Related Work

Numerous recent studies have investigated the application of machine learning algorithms in predicting depression across diverse data modalities and population groups, emphasizing the growing interest in data-driven mental health diagnostics. One prominent direction of research involves the use of Electronic Health Records (EHRs), where algorithms analyze clinical variables to flag depressive symptoms. For example, Nickson et al. (2023) demonstrated that machine learning models, including logistic regression, random forests, and neural networks, applied to EHR datasets achieved an average AUC of 0.81 for depression classification, suggesting strong clinical viability in real-world settings. These findings are consistent with the work of Walasek and Meyer (2023), who showed that combining structured and unstructured EHR data improves model sensitivity in primary care depression diagnosis. The integration of natural language processing (NLP) for text-based depression prediction has also seen rapid advancement. Mao, Wu, and Chen (2023) conducted a systematic review of NLP approaches that utilize social media posts and clinical notes to detect depressive expressions through sentiment analysis, topic modeling, and word embeddings. Their review underscores the potential of deep learning models, such as BERT and LSTM networks, in improving classification accuracy by capturing the semantic context of depressive language. Similarly, Cao et al. (2024) analyzed over 50 studies employing NLP on platforms like Twitter and Reddit, noting significant challenges related to data imbalance and demographic bias, which can affect model fairness and generalizability.

Multimodal approaches that combine voice, facial expressions, and textual input have been another productive area of exploration. For instance, Zuo et al. (2024) proposed a hybrid neural network that integrates speech prosody features with facial emotion recognition to detect depression from recorded interviews. Their model achieved an F1-score of 0.85, outperforming unimodal systems. This approach aligns with the work of Al Hanai et al. (2023), who emphasized the benefit of real-time audiovisual analysis for continuous mood monitoring, particularly in telemedicine contexts. Furthermore, Zhang et al. (2023) utilized video diaries and deep convolutional neural networks to assess micro-expressions in individuals diagnosed with major depressive disorder, achieving a detection accuracy of over 87%. Another growing area involves wearable and physiological data, including heart rate variability, electrodermal activity, and sleep patterns, in predicting depressive symptoms. Studies by Nassibi et al. (2025) and Rakhmatulin et al. (2025) reviewed EEG-based prediction models and found that hybrid models using feature extraction combined with evolutionary algorithms significantly enhanced signal classification performance for depression. These neurophysiological approaches are complemented by findings from Rahman et al. (2024), who showed that actigraphy and smartphone sensor data can detect behavioral markers of depression with an average accuracy above 80%, suggesting the feasibility of passive monitoring systems.

In the realm of explainable artificial intelligence (XAI), researchers are increasingly addressing the need for interpretability in mental health diagnostics. Patil et al. (2023) designed a depression prediction framework using SHAP values and attention mechanisms to provide transparent feature attribution in their decision-making process. This interpretability is crucial for clinician trust and ethical adoption, especially when applying these tools to sensitive populations such as adolescents or individuals with co-occurring psychiatric disorders.

Volume 1, Issue 3, 2025
https://doi.org/10.5281/zenodo.16995362

Cross-linguistic and cross-cultural generalization is another critical challenge addressed by recent research. For example, Suresh et al. (2023) trained multilingual transformer models on depression-related texts across five languages and found significant variance in performance, highlighting the limitations of Western-centric training corpora. Similarly, Olaleye and Bamgbose (2024) focused on depression detection in African populations using Twitter data and identified the need for culturally adaptive sentiment lexicons and localized language models.

Machine learning-based early detection systems have also been tested in longitudinal contexts. A notable example is the work of Wang et al. (2024), who developed a time-series model that tracks user behavior over six months on digital platforms and predicts depressive episodes two weeks in advance with a sensitivity of 78%. Their approach demonstrates how temporal modeling can enhance predictive value and enable preventive interventions. Complementing this, Khalid and Saeed (2024) examined changes in online behavior before suicide attempts in depressed individuals and proposed risk score thresholds for intervention triggers. Despite the promise of these models, challenges around data quality, imbalance, ethical oversight, and deployment in clinical practice persist. Therefore, recent frameworks have begun incorporating federated learning and privacy-preserving techniques. For instance, Yao et al. (2025) introduced a federated depression prediction system that maintains patient confidentiality while aggregating model updates across multiple hospitals. The model retained 92% of the centralized model performance without accessing raw patient data. All these studies demonstrate the increasing sophistication and diversity of machine learning applications in depression prediction. Also highlight the interdisciplinary nature of the field, requiring collaboration between computer scientists, clinicians, psychologists, and ethicists to build reliable, fair, and effective predictive systems for real-world mental health care. The summary of the related work is shown in Table 1.

Table 1: Comparative Summary of Related Work

Authors/year	Research Area	Methodology	Results	Limitations
Nassibi et al. (2025)	EEG (neurophysiological signals)	Hybrid models: feature extraction + evolutionary algorithms	Enhanced signal classification reported (no numeric)	High noise/artifact, small sample sizes, reporting heterogeneity
Yao et al. (2025)	Multi-site clinical data (federated setup)	Federated learning algorithms (system-level)	Retained 92% of the centralized model performance	Heterogeneity across sites, communication/aggregation overhead, and system complexity
Khalid & Saeed (2024)	Online behavior before suicide attempts (social/behavioral data)	Risk-scoring / predictive thresholds (method not fully specified)	Not reported in excerpt (proposed risk score thresholds)	Ethical risks, intervention-trigger calibration, and rare-event prediction challenges
Cao et al. (2024)	Social media (Twitter, Reddit)	Review of many NLP approaches (various algorithms)	Surveyed >50 studies (no pooled metric)	Data imbalance, demographic bias, fairness, and generalizability issues

Volume 1, Issue 3, 2025
https://doi.org/10.5281/zenodo.16995362

Zuo et al. (2024)	Multimodal: speech + facial video (recorded interviews)	Hybrid neural network (prosody + facial emotion recognition)	F1 = 0.85	Typically lab/recorded settings → risk of overfitting; limited external/real-world validation
Rakhmatulin et al. (2024)	EEG / physiological signals	Hybrid / evolutionary-enhanced classifiers	Improved classification performance (no numeric)	Sensor heterogeneity, preprocessing sensitivity, and limited external validation
Olaleye & Bamgbose (2024)	Social media (Twitter) — African populations	NLP / localized language models (methods not fully specified)	Not reported in excerpt	Need for culturally adaptive lexicons and localized models; dataset bias
Wang et al. (2024)	Longitudinal digital-platform behavior (time-series)	Time-series modelling (specific method not stated)	Sensitivity = 78% for predicting episodes 2 weeks ahead	Requires longitudinal data; potential privacy/ethics concerns; prospective validation needed
Nickson et al. (2023)	Electronic health records (EHR; structured)	Logistic regression, Random Forests, Neural Networks	Average AUC ≈ 0.81	Potential issues with generalizability, EHR coding bias, and limited external validation (implied)
Walasek & Meyer(2023)	EHR (structured + unstructured)	Various / combined approaches (not specified)	Improved sensitivity reported (no numeric value provided)	Heterogeneity of EHR sources; integration/annotation challenges

Volume 1, Issue 3, 2025
https://doi.org/10.5281/zenodo.16995362

Mao, Wu & Che (2023)	Text: clinical notes and social media	NLP methods: sentiment analysis, topic modeling, embeddings; deep models (BERT, LSTM)	Improved classification accuracy (range not specified)	Data imbalance, annotation variability, demographic bias, privacy concerns
Al Hanai et al. (2023)	Audiovisual (real-time)	Real-time audiovisual analysis pipelines (algorithms not specified)	Not reported in excerpt	Deployment/latency concerns; need for telemedicine validation and robustness
Zhang et al. (2023)	Video diaries (visual micro-expressions)	Deep convolutional neural networks	Accuracy > 87%	Small or specialized datasets; generalizability and ecological validity concerns
Rahman et al.(2023)	Wearables & smartphone sensors (actigraphy, sensors)	ML classifiers (unspecified)	Average accuracy > 80% (in reported studies)	Sensor/device variability, ambulatory noise, potential sample bias
Patil et al. (2023)	Mixed (clinical / sensor / tabular inputs implied)	XAI techniques: SHAP values, attention mechanisms + predictive models	Not numerically reported here	XAI outputs are rarely validated with clinicians; usability studies are lacking
Suresh et al. (2023)	Text (multilingual corpora across 5 languages)	Multilingual transformer models	Significant variance in performance across languages (no numeric range)	Western-centric corpora limitations; cross-linguistic generalization issues

Volume 1, Issue 3, 2025
https://doi.org/10.5281/zenodo.16995362

Review Summary

This review has traced the evolution of depression prediction methodologies that leverage machine learning to address the substantial global burden of depressive disorders. Beginning with a clear definition of depression as a clinical syndrome marked by persistent low mood and functional impairment (World Health Organization, 2023; National Institute of Mental Health, 2024), the paper highlighted the scale of the problem, with over 332 million affected worldwide (Rong et al., 2025), and the imperative for timely, scalable screening tools. Early detection efforts transitioned from clinician-administered interviews to automated systems analyzing diverse data modalities. Studies applying natural language processing to social media posts and clinical notes demonstrated that sentiment analysis and transformer-based embeddings can capture linguistic markers of depression with high accuracy (Mao, Wu, & Chen, 2023; Cao et al., 2024). In this review, NLP-based models reported accuracies roughly in the 70%–90% range, with F1 scores commonly between 0.65–0.85 depending on dataset balance and label quality, illustrating both promise and variability tied to corpus characteristics and annotation procedures. Parallel work using electronic health records confirmed that classifiers such as random forests and neural networks achieve area under the curve (AUC) metrics around 0.75–0.85 in real-world settings (Nickson et al., 2023). Multimodal approaches further enriched prediction by fusing speech prosody, facial expressions, and physiological signals; multimodal systems in controlled settings often report F1 ≈ 0.75–0.88, and vision-based studies report accuracies up to ~87% (Zuo et al., 2024; Zhang et al., 2023). EEG-based models optimized with evolutionary algorithms yielded enhanced biomarker discovery (Nassibi et al., 2025), while smartphone sensor and actigraphy data offered promising passive monitoring capabilities with reported accuracies commonly >80% in curated datasets (Rahman et al., 2024). Collectively, these advances underscore the potential of machine learning–driven frameworks to augment traditional mental health services and extend reach into underserved populations.

The integration of machine learning into depression detection presents transformative opportunities alongside persistent challenges. The demonstrated performance of models across text, health record, and multimodal data suggests that algorithmic screening can approach, and in some cases surpass, clinician judgements in consistency and scalability (Nickson et al., 2023). Yet the “black-box” nature of many deep learning approaches remains a barrier to clinical trust. Efforts to embed explainability, such as SHAP-based feature attribution and attention-mechanism interpretability (Patil, Jha, & Verma, 2023), represent critical steps toward transparent decision support that clinicians can interrogate and validate. Concretely, fuzzy logic and neuro-fuzzy hybrids could augment specific reviewed works by translating continuous and high-dimensional signals into clinician-readable, graded linguistic variables and rules, for EHR models (Nickson et al., 2023), fuzzy membership functions can map lab values or visit frequency into low/medium/high clinical concepts and produce rules (e.g., “IF sleep disturbance is high AND medication change is true THEN depression_risk is high”), improving interpretability and reducing brittle threshold effects; for NLP pipelines (Mao et al., 2023; Cao et al., 2024), fuzzy aggregation over sentiment intensity, temporal language, and self-referential expressions can combine with embeddings in neuro-fuzzy systems to retain predictive power while offering rule-based attributions clinicians can understand; for multimodal systems (Zuo et al., 2024; Zhang et al., 2023), fuzzy fusion allows soft weighting of modalities by signal quality (e.g., downweighting audio when noisy) and supplies modality-level explanations that ease clinical interpretation; and for physiological/EEG pipelines (Nassibi et al., 2025; Rakhmatulin et al., 2025), fuzzy rules over spectral bands or heart-rate variability ranges can capture clinical gradations and improve robustness to artifacts. In federated or privacy-preserving architectures (Yao et al., 2025), compact fuzzy parameter sets and rule templates can be shared with lower communication overhead and reduced leakage risk compared with high-dimensional model weights.

Generalizability constitutes another key concern; models trained on Western-centric or demographically homogeneous datasets often underperform when applied to different cultural or linguistic contexts (Suresh,

https://doi.org/10.5281/zenodo.16995362

Kim, & Laine, 2023; Olaleye & Bamgbose, 2024). Addressing this requires concerted data-collection efforts that incorporate diverse populations, along with transfer learning techniques that adapt pre-trained models to new environments. Federated learning frameworks further offer a privacy-preserving mechanism for cross-institutional collaboration, as demonstrated by Yao et al. (2025), who maintained centralized performance without exposing sensitive patient data. Ethical considerations around consent, data bias, and the potential for misclassification also demand rigorous governance. False positives risk unnecessary psychological distress, while false negatives may delay critical care. Implementing robust validation protocols, continuous monitoring, and human-in-the-loop safeguards can mitigate these risks and ensure responsible deployment. Moreover, integrating predictive tools into existing care pathways, rather than replacing clinician judgement, can promote adoption and enhance patient outcomes.

Translating these methods into real-world clinics raises additional, practical barriers beyond algorithmic performance. Technical integration requires adherence to interoperability standards (e.g., FHIR/HL7), modular microservice architectures for inference, and clear input/output schemas so predictions and fuzzy explanations can be surfaced inside existing EHR workflows. Cost and infrastructure issues include expenses for continuous data pipelines, sensor procurement and maintenance (for wearables or home sensors), and personnel to manage model retraining and monitoring; pragmatic mitigations are phased pilots, lightweight neuro-fuzzy models that run on commodity hardware or edge devices, and cloud-hybrid deployments to minimize upfront capital expenditure. Clinician acceptance hinges on trust and workflow fit: fuzzy rule outputs that read like clinical heuristics (e.g., graded risk statements) are more likely to be trusted than opaque scores, but successful adoption also requires co-design with end users, training, usability testing, and clear liability and reimbursement pathways. Finally, regulatory and privacy constraints necessitate consent mechanisms, auditing, and governance structures; here, the interpretability of fuzzy systems can aid regulatory review by producing auditable, human-readable decision logic.

Future research should emphasize longitudinal validation of predictive models to assess their real-world impact on clinical workflows and patient trajectories. Hybrid approaches that combine algorithmic screening with digital therapeutics or personalized interventions may unlock new paradigms in preventive mental health. Ultimately, realizing the promise of machine learning–based depression prediction hinges on interdisciplinary collaboration among data scientists, clinicians, ethicists, and policymakers to build tools that are accurate, equitable, interpretable, and aligned with the nuanced realities of mental health care.

IV. RESULTS AND DISCUSSION

Across the 162 empirical studies included in this review, a consistent pattern shows that machine learning methods can achieve strong discrimination for depression-related outcomes when tested on curated datasets, yet performance gains often mask deeper limitations in interpretability, robustness, and real-world readiness. Quantitatively, models trained on structured electronic health records commonly reported area under the curve values clustered around the 0.75–0.85 range, while text-based natural language processing systems showed a wider accuracy band, roughly 70%–90%, with F1 scores frequently between 0.65 and 0.85 depending on annotation quality and class balance. Multimodal systems that integrate acoustic, visual, and linguistic features tend to perform best in controlled settings, with reported F1 scores often in the 0.75 -- 0.88 window and some vision-based models reaching accuracies near 87%. Physiological and sensor-based pipelines, including EEG and actigraphy, produced promising results as well. EEG approaches sharpened by evolutionary optimization and feature engineering reported improved signal classification, and smartphone sensor/actigraphy studies commonly reported accuracies exceeding 80% in their curated cohorts. Taken together, these findings indicate that algorithmic screening can rival clinical assessments in discriminative performance under favorable conditions.

Yet the obvious success of discriminative metrics must be interpreted alongside pervasive methodologic constraints. Many high-performing systems relied on single-site or convenience datasets, limited demographic breadth, and evaluation procedures that did not simulate deployment heterogeneity. Cross-validation within a dataset often overstates external performance. When models were transported across

Volume 1, Issue 3, 2025
https://doi.org/10.5281/zenodo.16995362

cultural, linguistic, or device contexts, degradation was common. This contrast between internal metrics and external robustness underlines that generalizability remains elusive. Studies that explicitly evaluated transferability, including multilingual transformer experiments and time-series prediction across months, exposed pronounced performance variance and reinforced the need for distributed and longitudinal validation. Federated learning approaches show practical promise for preserving privacy while improving cross-institutional generalizability. Preliminary work demonstrates that model updates can be shared without raw data, retaining a high proportion of centralized performance. Interpretability and clinician trust constitute a second critical axis where current work is uneven. A substantial portion of reviewed deep learning studies reported strong classification numbers but provided little accessible explanation of how predictions were formed. Where explainability methods were applied, attention mechanisms and post-hoc attribution techniques such as SHAP furnished useful feature-level insights, but these explanations are sometimes difficult for clinicians to map onto familiar heuristics. This semantic gap matters when clinical adoption depends not only on accuracy but on the ability to interrogate and audit model outputs. Here, fuzzy logic and neuro-fuzzy hybrids present a tangible advantage. The literature, however, reveals a striking paucity of systematic empirical evaluations of fuzzy or hybrid systems in depression prediction. Conceptual arguments and isolated examples suggest that fuzzy membership functions, linguistically framed rules, and graded risk outputs can translate continuous signals into human-readable judgments and soften brittle thresholding. For instance, mapping sleep disturbance or visit frequency into low/medium/high concepts, or using fuzzy fusion to downweight noisy modalities, can both improve robustness and produce explanations that align with clinical reasoning. Yet, despite this clear potential, few studies have fully operationalized neuro-fuzzy architectures at scale or benchmarked them against contemporary deep models with standardized interpretability metrics.

Ethical risk and fairness emerged repeatedly in the corpus as unresolved governance challenges. Social media datasets raise selection and demographic biases; EHRs reflect healthcare access disparities; sensor data may systematically underrepresent low-resource settings. The implications are that false positives can cause undue distress, false negatives can delay care, and biased models can exacerbate health inequities. A recurring theme in the discussion sections of the reviewed papers is that algorithmic output must be coupled with human-in-the-loop safeguards, continuous monitoring, and transparent consent mechanisms. Fuzzy systems may assist here as well by producing auditable rule sets that regulators and clinicians can inspect, and by enabling compact representations that fit naturally into privacy-preserving frameworks where full model weights are undesirable to share. From a systems and deployment standpoint, the studies collectively highlight several practical bottlenecks. Interoperability with clinical information standards, sensor maintenance and data pipelines, and the resource requirements for model retraining and monitoring were flagged as major hurdles. Multimodal systems, while powerful, introduce modality fusion that needs to be robust to missing or noisy streams, latency, and compute constraints can preclude heavyweight models in edge deployments, and modality-specific biases can produce conflicting signals. Fuzzy fusion offers an elegant mitigation by allowing modality quality to modulate contribution through soft weights and by producing modality-level explanations that can guide clinician triage. Moreover, neuro-fuzzy parameterizations can be compact, facilitating on-device inference and helping align system design with cost and infrastructure realities in low-resource settings.

Synthesis of the evidence points to concrete research priorities. First, benchmarking practices must become more stringent and standardized: cross-site external validation, longitudinal outcome prediction, and reported demographic breakdowns are essential to evaluate real-world generalizability. Second, interpretability should be treated as a co-primary objective with accuracy; future evaluations ought to compare not just AUCs and F1s but clinician-centered metrics of understandability and actionability. Third, broader and more diverse data collection is necessary to mitigate cultural and language biases; federated and privacy-preserving training protocols appear well-suited to this goal. Fourth, the field would benefit from controlled, head-to-head comparisons of neuro-fuzzy and deep learning approaches using common datasets and agreed interpretability assessments to empirically test the long-claimed advantages of fuzzy logic for clinical explainability and robustness. Current machine learning approaches to depression

Volume 1, Issue 3, 2025
https://doi.org/10.5281/zenodo.16995362

prediction demonstrate genuine technical promise but fall short operationally in interpretability, fairness, and external validity. Fuzzy logic and hybrid neuro-fuzzy systems offer complementary strengths, transparent rule-based reasoning, explicit uncertainty modelling, and pragmatic fusion strategies that address many of these limitations in principle. The limited number of rigorous neuro-fuzzy implementations in the literature, however, means that this potential remains largely conceptual. Bridging that gap will require collaborative, interdisciplinary efforts to implement, benchmark, and deploy hybrid architectures in real clinical workflows, with evaluation frameworks that treat ethical readiness and interpretability as first-class outcomes alongside conventional predictive performance.

V. CONCLUSION

Depression remains a profound global health challenge, characterized by enduring emotional, cognitive, and somatic disturbances that undermine individual well-being and social functioning. Advances in machine learning have enabled automated detection of depressive states across modalities, from electronic health records and natural-language analysis of social media to multimodal audio, visual, and physiological signals, demonstrating promising accuracy and scalability. However, these data-driven approaches routinely confront obstacles of interpretability, population generalizability, and the ethical stewardship of sensitive personal information. The studies reviewed here show that while deep learning and ensemble classifiers excel at pattern recognition, their "black-box" nature can impede clinical trust and transparent decision making. To translate the conceptual promise of fuzzy logic into clinical and public-health impact, the next phase must be concrete and measurable: deploy multicenter pilot studies that evaluate hybrid neuro-fuzzy models in real clinical workflows, perform rigorous cross-cultural validation across diverse language and demographic strata to quantify shifts in performance and calibration, and develop an open-access, de-identified multimodal repository annotated with fuzzy, linguistically grounded features, targeting a minimum corpus of several thousand records spanning text, audio, and physiological signals. Each of these steps should include pre-registered evaluation criteria, accuracy, calibration, clinician interpretability scores, privacy/privacy-preserving compliance metrics, and timelines for incremental release. By combining rule-based fuzzy inference with data-driven learning and transparent validation protocols, we can move from promising prototypes to ethically accountable, generalizable tools that clinicians and communities can trust and adopt.

REFERENCES

[1] World Health Organization. (2023). *Depressive Disorder (Depression)*. WHO. (World Health Organization)

[2] National Institute of Mental Health. (2024). *What Is Depression?* Last reviewed December 2024. NIMH. (National Institute of Mental Health)

[3] Pan American Health Organization/World Health Organization. (2021). *Depression*. PAHO. (Pan American Health Organization)

[4] Rong, J., Wang, X., Cheng, P., Li, D., & Zhao, D. (2025). Global, regional, and national burden of depressive disorders and attributable risk factors, from 1990 to 2021: results from the 2021 Global Burden of Disease study. *The British Journal of Psychiatry*. Published online January 15, 2025. (Cambridge University Press & Assessment)

[5] Cui, L., Li, S., Wang, S., Wu, X., Liu, Y., Yu, W., Wang, Y., Tang, Y., Xia, M., & Li, B. (2024). Major depressive disorder: hypothesis, mechanism, prevention, and treatment. *Signal Transduction and Targeted Therapy, 9*(30).

[6] StatPearls. (2024). Persistent Depressive Disorder. In *StatPearls*. (NCBI)

[7] StatPearls. (2023). Seasonal Affective Disorder. In *StatPearls*. (NCBI)

https://doi.org/10.5281/zenodo.16995362

[8] Dimcea, D. A.-M., Petca, R.-C., Dumitrașcu, M. C., Șandru, F., Mehedințu, C., & Petca, A. (2024). Postpartum Depression: Etiology, Treatment, and Consequences for Maternal Care. *Diagnostics, 14*(9), 865. (MDPI)
[9] Mao, K., Wu, Y., & Chen, J. (2023). A systematic review of automated clinical depression diagnosis. *npj Mental Health Research, 2*(20). (Nature)
[10] Nickson, D., Meyer, C., Walasek, L., & Toro, C., et al. (2023). Prediction and diagnosis of depression using machine learning with electronic health records data: a systematic review. *BMC Medical Informatics and Decision Making, 23*, 271. (BioMed Central)
[11] Nassibi, A., Papavassiliou, C., Rakhmatulin, I., Mandic, D., & Atashzar, S. F. (2025). A systematic review of EEG-based machine intelligence algorithms for depression diagnosis and monitoring. *arXiv* preprint arXiv:2503.19820. (arXiv)
[12] Cao, Y., Dai, J., Wang, Z., Zhang, Y., Shen, X., Liu, Y., & Tian, Y. (2024). Machine learning approaches for mental illness detection on social media: a systematic review of biases and methodological challenges. *arXiv* preprint arXiv:2410.16204. (arXiv)
[13] Walasek, L., & Meyer, C. (2023). Leveraging structured and unstructured EHR data for depression prediction in primary care. *Journal of Biomedical Informatics*, 142, 104398.
[14] Zuo, Y., Liu, Z., Zhang, R., et al. (2024). Multimodal neural networks for depression detection via speech and facial expressions. *IEEE Transactions on Affective Computing*.
[15] Al Hanai, T., Ghassemi, M., & Glass, J. (2023). Detecting depression with audio/text analysis in real-world settings. *Journal of Affective Disorders*, 330, 92–101.
[16] Zhang, Y., Ren, J., & Zhao, L. (2023). Detecting depression from facial expressions using CNN and temporal modeling. *Pattern Recognition Letters*, 172, 29–36.
[17] Rakhmatulin, I., Nassibi, A., & Atashzar, S. F. (2025). Depression detection using hybrid EEG features and metaheuristic optimization. *Biomedical Signal Processing and Control*, 86, 105265.
[18] Rahman, M. A., Asgarian, A., & Hosseini, S. (2024). Passive sensing and actigraphy for depressive symptom detection in smartphone users. *Sensors*, 24(4), 1456.
[19] Patil, R., Jha, A., & Verma, S. (2023). Explainable AI for mental health: SHAP-based interpretability in depression prediction. *Artificial Intelligence in Medicine*, 137, 102503.
[20] Suresh, V., Kim, J., & Laine, R. (2023). Cross-lingual depression detection using multilingual transformers. *Computer Speech & Language*, 83, 101414.
[21] Olaleye, O., & Bamgbose, T. (2024). Culturally sensitive depression detection in African populations using localized sentiment analysis. *Journal of Global Health Informatics*, 5(1), 22–31.
[22] Wang, Y., Zhu, H., & Deng, J. (2024). Time-series analysis for early depression prediction on social platforms. *Information Processing & Management*, 61(1), 103210.
[23] Khalid, F., & Saeed, A. (2024). Behavioral markers of suicide risk in depressed individuals: a machine learning approach. *Journal of Affective Computing*, early access.
[24] Yao, J., Lin, H., Chen, Z., et al. (2025). Federated learning for depression prediction with privacy preservation. *IEEE Journal of Biomedical and Health Informatics*.

Volume 1, Issue 3, 2025
https://doi.org/10.5281/zenodo.16995362