International Journal of Pharmaceutical research and Applications
Volume 11, Issue 5, Sep.-Oct.2026,pp:1048-1054 www.ijprajournal.com ISSN:2456-4494

Privacy-Preserving Machine Learning for Mental Health Risk Prediction

[1]Sonal Shahi, [2]Pradeep Yadav
[1,2]*Department of Computer Science and Engineering, ITM Gwalior, Madhya Pradesh, India*

Date of Submission: 11-10-2026 Date of Acceptance: 22-10-2026

Abstract—Machine learning can support earlier identification of mental-health risk from electronic health records, speech, social-media text, wearable sensors, and other digital traces, but centralized data collection creates substantial privacy and governance concerns. This paper presents a literature-grounded analysis of privacy-preserving machine learning (PPML) for anxiety and depression risk prediction, focusing on federated learning, differential privacy, secure aggregation, homomorphic encryption, and trusted execution environments. The study compares the privacy guarantees, utility implications, computational requirements, and deployment constraints of these approaches and examines recent evidence from mental-health applications. Particular attention is given to federated depression detection from multilingual social-media text, privacy-preserving speech-based depression diagnosis, and federated classification of major depressive disorder from EEG. The analysis shows that federated learning can reduce the need to centralize raw sensitive data, but it does not eliminate privacy risks: model updates may still leak information and can be exposed to poisoning or inference attacks. Differential privacy provides formal protection but introduces a measurable privacy–utility trade-off. The paper proposes a layered deployment framework combining local training, secure aggregation, differential privacy, robustness checks, and human-centred governance. The study concludes that privacy protection should be evaluated jointly with predictive performance, calibration, fairness, robustness, communication cost, and clinical utility rather than treated as an independent feature.

Keywords—privacy-preserving machine learning, federated learning, differential privacy, mental health, depression prediction

I. INTRODUCTION

Mental-health prediction increasingly uses computational analysis of information generated during routine care and everyday digital activity. Electronic health records can contain longitudinal diagnoses and clinical observations; speech and text can provide behavioral and linguistic signals; and smartphones or wearables can capture patterns associated with sleep, activity, mobility, and social interaction. Such information can be valuable for research and early-risk screening, but mental-health data are highly sensitive and may expose individuals to stigma, discrimination, or unwanted disclosure if handled inappropriately.

Conventional centralized machine learning typically transfers data to a common repository for training. This architecture simplifies model development but increases the amount of sensitive information that must be collected, stored, governed, and potentially shared. Federated learning (FL) changes the data-flow model by keeping training data at participating clients and exchanging model information instead of raw records [1], [2]. However, FL should not be described as synonymous with complete privacy. Gradients or model updates can still reveal information, and distributed systems introduce risks including poisoning, inference, insecure aggregation, and heterogeneous client behavior [1], [2].

Privacy-preserving machine learning (PPML) therefore requires a combination of technical and organizational controls. Differential privacy (DP) limits the influence of individual records through controlled randomization [3]. Secure aggregation allows a server to obtain an aggregate update without directly seeing each client's individual contribution [4]. Homomorphic encryption (HE) can enable computation on encrypted values, while trusted execution environments (TEEs) seek to protect computation within hardware-supported isolated regions. These mechanisms have different costs and threat models.

This paper develops a comparative, literature-based analysis of PPML for mental-health risk prediction. It asks: (1) which privacy mechanisms are most suitable for different mental-health data modalities; (2) how privacy protection affects predictive utility and system efficiency; and (3) what methodological and governance safeguards

International Journal of Pharmaceutical research and Applications
Volume 11, Issue 5, Sep.-Oct.2026,pp:1048-1054 www.ijprajournal.com ISSN:2456-4494

are needed before deployment. The paper does not claim a new clinical model or report original patient experiments; reported metrics are explicitly attributed to the cited studies.

II. BACKGROUND AND RELATED WORK

A. Mental-Health Prediction and Sensitive Data

Machine learning has been studied for mental-health prediction using clinical records, social-media language, speech, images, and passive sensing. Shatte et al. reviewed 300 papers covering ML and big-data applications in mental health and emphasized substantial variation in methods and data sources [5]. Nemesure et al. demonstrated predictive modeling of depression and anxiety using EHR data [6]. Social-media research has also shown that linguistic and behavioral features can be associated with depression-related outcomes [7], [8]. These findings establish the predictive potential of digital data but also highlight why privacy protection is necessary.

The privacy challenge is particularly pronounced when the input itself is intimate. Speech contains biometric and behavioral information; social-media posts can reveal relationships and personal experiences; EHRs combine clinical, demographic, and administrative information. A privacy-preserving design must therefore consider not only database security but also what can be inferred from trained models and transmitted updates.

B. Federated Learning

FL was introduced as a decentralized approach in which participating clients train locally and a coordinating server aggregates model updates [1]. FedAvg is a foundational aggregation method that computes a weighted average of client model parameters [1]. In healthcare, FL is attractive because institutions can collaborate without creating a single repository of raw patient records. Surveys have documented applications in medical imaging, disease prediction, EHR analysis, and IoMT [2], [9].

Mental-health studies provide more specific evidence. Privacy-preserving speech-based depression diagnosis demonstrated that federated models could be trained on the DAIC-WOZ dataset while retaining speech data on clients and incorporating norm bounding, differential privacy, and secure aggregation as defenses [10]. A 2024 study evaluated federated depression detection using multilingual social-media data in English, Arabic, Spanish, Russian, and Korean, finding that FL could perform strongly while keeping client data decentralized; it also showed that non-IID data and severe quantity imbalance affect performance [11].

A 2024 EEG study investigated federated classification of major depressive disorder using LSTM, bidirectional LSTM, GRU, and 1-D CNN models under non-IID conditions [12]. These studies indicate that FL is feasible for mental-health applications, but the evidence remains dataset-specific and should not be generalized to clinical effectiveness without external and prospective validation.

C. Differential Privacy

DP provides a mathematical privacy guarantee by bounding how much the output of an algorithm can change when one individual's record is added or removed. In ML, common implementations clip per-example gradients and add calibrated noise during training [3]. The privacy budget, commonly represented by epsilon (ε), controls the strength of the guarantee; stronger privacy generally requires more noise and may reduce predictive utility.

DP can complement FL rather than replace it. FL keeps raw records decentralized, while DP reduces the information that can be inferred from updates or outputs. Recent mental-health work illustrates the trade-off. FedMental evaluated federated and differentially private federated learning for depression detection on social-media data and reported that ordinary FL approached centralized performance, whereas DP introduced substantial utility loss under the evaluated settings [13]. This result reinforces the need to report both privacy parameters and predictive metrics.

D. Other Privacy-Enhancing Technologies

Secure multiparty computation and secure aggregation can prevent a central coordinator from inspecting individual client updates. Homomorphic encryption permits selected computations over encrypted values but can increase computation and communication requirements. TEEs can protect sensitive operations through hardware isolation, although their security depends on hardware, implementation, and trust assumptions. Recent healthcare surveys emphasize that these technologies can be combined rather than treated as mutually exclusive choices [2], [14].

III. METHODOLOGY

This paper uses a structured narrative synthesis of peer-reviewed research and methodological literature on PPML, with emphasis on mental-health

International Journal of Pharmaceutical research and Applications
Volume 11, Issue 5, Sep.-Oct.2026,pp:1048-1054 www.ijprajournal.com ISSN:2456-4494

prediction and healthcare FL. Sources were selected for direct relevance to privacy-preserving training, mental-health prediction, or foundational privacy mechanisms. Priority was given to peer-reviewed journal articles, conference papers, systematic reviews, and established methodological papers. Recent studies from 2024–2026 were included where verifiable.

Each study was examined across six dimensions: data modality, prediction task, privacy mechanism, learning architecture, evaluation evidence, and reported limitations. Performance values are retained only as study-specific results. Because datasets, labels, populations, prediction horizons, and validation protocols differ, the paper does not construct a cross-study accuracy leaderboard.

The analytical criteria are predictive utility, privacy strength, computational overhead, communication burden, robustness, interpretability, fairness, and deployment feasibility. This framework is intended to help researchers choose a privacy architecture according to threat model and operational context rather than selecting FL or DP solely because it is considered privacy-preserving.

IV. PROPOSED PRIVACY-PRESERVING FRAMEWORK

A practical PPML architecture for mental-health risk prediction can be organized into five layers. The design is conceptual and synthesizes mechanisms reported in the literature rather than representing a validated clinical system.

Fig. 1. Conceptual five-layer architecture: (1) local data and preprocessing, (2) local model training, (3) privacy protection, (4) secure aggregation and global optimization, and (5) governance, evaluation, and human review.

A. Local Data Layer

Data remain within the originating institution or device whenever feasible. Local preprocessing includes normalization, missing-value handling, label verification, and feature extraction. The system should minimize collection and avoid transferring raw mental-health records to a central coordinator.

B. Local Learning Layer

Each client trains a shared model using its local data. Candidate models can include logistic regression or tree-based methods for structured features and neural models for text, speech, EEG, or longitudinal records. A simple baseline should always be retained so that added model complexity can be justified.

C. Privacy Protection Layer

The third layer combines mechanisms according to the threat model. Gradient clipping and DP can reduce information leakage; secure aggregation can hide individual client updates; encryption can protect data in transit or during computation. Combining controls increases protection but also adds computational and operational cost.

D. Aggregation Layer

The coordinator receives protected updates and computes a global model using an aggregation rule such as FedAvg [1]. Robust aggregation can be considered where malicious or unreliable clients are plausible. Client selection, dropout, communication frequency, and non-IID distributions should be logged because they affect convergence and fairness.

E. Governance and Evaluation Layer

The final layer covers access control, audit logs, privacy accounting, subgroup evaluation, model monitoring, human review, and incident response. A high-risk prediction should be treated as decision support rather than an autonomous diagnosis. Governance should specify who can access outputs, how long information is retained, how errors are corrected, and how individuals can exercise appropriate rights.

TABLE I. COMPARISON OF PRIVACY-PRESERVING APPROACHES

Approach	Data exposure	Privacy level*	Overhead	Main issue
Centralized ML	Raw data centralized	Low–medium	Simple	High central exposure
FL	Raw data local; updates shared	Medium	Medium	Update leakage possible
FL + Secure Aggregation	Raw data local; aggregate updates	Medium–high	Medium	Protects individual updates
FL + DP	Local data + noisy updates	High	Medium–high	Formal privacy; utility loss
FL + DP + Secure Agg.	Layered protection	High	High	Stronger protection; complexity
HE-assisted FL	Encrypted	Very high	Very high	Strong

International Journal of Pharmaceutical research and Applications
Volume 11, Issue 5, Sep.-Oct.2026,pp:1048-1054 www.ijprajournal.com ISSN:2456-4494

	computation/updates			confidentiality; overhead
TEE-assisted FL	Hardware-isolated computation	Medium–high	High	Hardware trust assumptions

*Privacy level is a qualitative architectural assessment, not a standardized numerical score.

V. COMPARATIVE EVIDENCE

Table II summarizes representative mental-health PPML evidence. The values are not directly comparable because the studies use different modalities and experimental settings.

TABLE II. REPRESENTATIVE PPML EVIDENCE IN MENTAL-HEALTH APPLICATIONS

Study	Data	Privacy method	Reported result	Caution
Speech depression [10]	DAIC-WOZ speech	FL + DP + secure aggregation	High performance with limited utility loss vs centralized in evaluated settings	Dataset-specific; simulated clients
Multilingual depression [11]	5-language social media	FL	Strong performance in many evaluated settings	Non-IID and quantity imbalance
MDD from EEG [12]	3-channel EEG	FL + DL	Evaluated LSTM, BiLSTM, GRU, 1D-CNN	Non-IID distributed training
FedMental [13]	X/Twitter + Reddit	FL; DP-FL	Central F1 85.63; best FL F1 83.16; DP-FL showed large drops in tested settings	Privacy–utility trade-off
DP-CARE [15]	Dreaddit text	DP fine-tuning	F1 78.08%; recall 88.67% at ε≈3	Stress detection, not depression

The speech study is particularly relevant because it combines federated learning with multiple defenses rather than assuming that decentralization alone guarantees privacy [10]. The multilingual depression study adds an important fairness and generalizability dimension: language and client distributions create non-IID conditions that can change model behavior [11].

The 2026 FedMental results provide a useful quantitative illustration of the privacy–utility tension. Ordinary FL produced a best reported F1 of 83.16 compared with 85.63 for centralized training, while the differentially private setting produced substantially larger performance losses in the tested configurations [13]. These results should be read as evidence of a trade-off, not as a universal estimate of DP's effect.

DP-CARE, although focused on stress detection rather than anxiety or depression, demonstrates a complementary approach: differential privacy can be applied during lightweight classifier training on a frozen mental-health language encoder [15]. Its reported F1 and recall show that privacy-aware training can remain useful, but the task and dataset differ from depression prediction and therefore should not be conflated.

VI. THREAT MODEL AND PRIVACY RISKS

A. Inference and Reconstruction

Model updates may reveal information about training examples. An attacker with access to gradients, updates, or model outputs may attempt membership inference or reconstruction. Consequently, keeping raw data on clients is insufficient as a complete threat model. Secure aggregation and DP can reduce different parts of this risk, but their protection assumptions should be stated explicitly.

B. Poisoning and Byzantine Behaviour

Distributed training expands the number of participants that can influence a global model. A malicious or compromised client may submit manipulated updates. Robust aggregation, anomaly detection, client authentication, and audit mechanisms can reduce this risk. These controls

International Journal of Pharmaceutical research and Applications
Volume 11, Issue 5, Sep.-Oct.2026,pp:1048-1054 www.ijprajournal.com ISSN:2456-4494

may conflict with privacy because detecting malicious behavior sometimes requires observing information that privacy mechanisms intentionally conceal.

C. Re-identification and Linkage

Even when direct identifiers are removed, mental-health data can remain identifiable through combinations of demographic, temporal, linguistic, or behavioral attributes. Data minimization and local processing should therefore be complemented by access controls and privacy risk assessment.

D. Privacy Budget and Composition

Repeated DP training or release of multiple model outputs consumes privacy budget. Privacy accounting should therefore cover the entire training process rather than reporting a single epsilon value without explaining the number of iterations, sampling assumptions, clipping norm, and noise mechanism.

VII. PERFORMANCE AND IMPLEMENTATION TRADE-OFFS

TABLE III. RECOMMENDED EVALUATION DIMENSIONS

Dimension	Example metrics	Evaluation focus
Predictive utility	AUC, AUPRC, F1, sensitivity, specificity	Compare with centralized and local baselines
Calibration	Brier score, calibration curve	Essential when output is a risk probability
Privacy	ε, δ, threat model	Report mechanism and accounting
Communication	Rounds, bytes/client, dropout	Measure end-to-end federation cost
Computation	Training time, memory, energy	Measure client and server separately
Fairness	Subgroup metrics	Check demographic/language variation
Robustness	Poisoning/inference tests	Evaluate under realistic attacks
Clinical utility	Alert burden, workflow outcomes	Prospective human evaluation

A PPML system can be technically private but operationally unsuitable if it requires excessive communication, long training times, or frequent client availability. Conversely, a highly accurate model may be unacceptable if it exposes sensitive updates or produces systematically poorer predictions for particular groups. Evaluation should therefore treat privacy, utility, efficiency, and fairness as a multi-objective problem.

For mental-health risk prediction, calibration deserves special emphasis. A model that estimates a 0.80 risk should produce approximately that frequency of the outcome in comparable cases if the probability is well calibrated. High discrimination alone does not guarantee useful probabilities. Prospective evaluation should also measure whether model alerts improve access to appropriate support without generating unnecessary burden or stigma.

VIII. ETHICAL AND GOVERNANCE CONSIDERATIONS

Privacy-preserving technology does not remove ethical responsibilities. Appropriate consent or another valid governance basis should be established for collection and secondary use. Participants should understand, in accessible language, what data are processed, whether computation occurs locally, what information is transmitted, and what the resulting prediction is used for.

Models should not be used to make unsupported inferences about a person's mental state or to impose consequential decisions without appropriate human review. False positives may create anxiety or stigma, while false negatives may delay support. These risks make model outputs unsuitable as stand-alone diagnoses.

Fairness must be evaluated across relevant demographic, linguistic, and socioeconomic groups. The multilingual FL evidence shows why client and language heterogeneity matters [11]. Privacy mechanisms can also affect groups differently if some clients have smaller datasets or different label distributions. Thus, privacy and fairness should be assessed jointly.

Governance should define data ownership, access permissions, retention, auditability, incident response, model-update procedures, and mechanisms for correcting erroneous information. A privacy-preserving architecture should be designed around

International Journal of Pharmaceutical research and Applications
Volume 11, Issue 5, Sep.-Oct.2026,pp:1048-1054 www.ijprajournal.com ISSN:2456-4494

the minimum necessary data flow rather than simply adding DP or FL after model development.

IX. LIMITATIONS AND FUTURE RESEARCH

The principal limitation of this paper is that it is a literature-based comparative study rather than an original clinical experiment. The cited mental-health studies use different datasets, tasks, languages, modalities, privacy parameters, and validation designs. Reported performance therefore cannot be interpreted as a common benchmark.

Future research should develop larger and more diverse federated mental-health benchmarks with transparent labels and reproducible privacy accounting. Longitudinal evaluation is needed to determine whether privacy-preserving models remain calibrated as populations, platforms, and clinical practices change. External validation across institutions and cultures is especially important.

Hybrid PPML architectures deserve further study. FL combined with secure aggregation and DP may provide stronger protection than any single mechanism, but the cumulative cost can be substantial. Efficient client selection, adaptive privacy budgets, communication compression, and on-device inference may reduce overhead. Research should also investigate robustness against model inversion, membership inference, poisoning, and collusion.

Finally, emerging foundation models and large language models create new privacy questions because memorization and sensitive-language representations can occur at scale. Lightweight private fine-tuning, parameter-efficient adaptation, privacy auditing, and provenance controls are promising directions, but mental-health deployment should remain subject to rigorous validation and human-centred governance.

X. CONCLUSION

Privacy-preserving machine learning provides a promising route for collaborative mental-health risk prediction without requiring routine centralization of raw sensitive data. Federated learning is particularly attractive because it changes the data-sharing architecture, while differential privacy, secure aggregation, encryption, and trusted execution can address additional threat surfaces. Evidence from speech, social-media, and EEG applications indicates technical feasibility, but also demonstrates that non-IID data, privacy budgets, and system constraints can affect predictive performance.

The central conclusion is that privacy cannot be reduced to a single technology or a single metric. A trustworthy system should report predictive discrimination and calibration together with privacy guarantees, communication and computational cost, robustness, fairness, and clinical utility. The proposed five-layer framework provides a practical structure for combining these considerations. Before deployment, models should undergo external and prospective validation and should support—not replace—appropriate human assessment.

REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. 20th Int. Conf. Artificial Intelligence and Statistics (AISTATS), 2017, pp. 1273–1282.

[2] M. Ali, F. Naeem, M. Tariq, and G. Kaddoum, "Federated learning for privacy preservation in smart healthcare systems: A comprehensive survey," IEEE J. Biomed. Health Inform., vol. 27, no. 2, pp. 778–789, 2023, doi: 10.1109/JBHI.2022.3181823.

[3] C. Dwork, A. Roth, "The algorithmic foundations of differential privacy," Found. Trends Theor. Comput. Sci., vol. 9, nos. 3–4, pp. 211–407, 2014, doi: 10.1561/0400000042.

[4] K. Bonawitz et al., "Practical secure aggregation for privacy-preserving machine learning," in Proc. ACM SIGSAC Conf. Computer and Communications Security, 2017, pp. 1175–1191, doi: 10.1145/3133956.3133982.

[5] A. B. R. Shatte, D. M. Hutchinson, and S. J. Teague, "Machine learning in mental health: A scoping review of methods and applications," Psychol. Med., vol. 49, no. 9, pp. 1426–1448, 2019, doi: 10.1017/S0033291719000151.

[6] M. D. Nemesure, M. V. Heinz, R. Huang, and N. C. Jacobson, "Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence," Sci. Rep., vol. 11, art. no. 1980, 2021, doi: 10.1038/s41598-021-81368-4.

[7] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in Proc. 7th Int. AAAI Conf. Weblogs and Social Media, 2013, pp. 128–137, doi: 10.1609/icwsm.v7i1.14432.

[8] J. C. Eichstaedt et al., "Facebook language predicts depression in medical records," Proc. Natl. Acad. Sci. USA, vol. 115, no. 44, pp. 11203–11208, 2018, doi: 10.1073/pnas.1802331115.

[9] A. Rieke et al., "The future of digital health with federated learning," NPJ Digit. Med., vol. 3, art. no. 119, 2020, doi: 10.1038/s41746-020-00323-1.
[10] Y. H. S. et al., "Privacy-preserving speech-based depression diagnosis via federated learning," 2022. [PubMed record; study using DAIC-WOZ and FL defenses.]
[11] "Federated learning for privacy-preserving depression detection with multilingual language models in social media posts," Patterns, vol. 5, no. 7, art. no. 100990, 2024, doi: 10.1016/j.patter.2024.100990.
[12] "Privacy Preserving Collaboratively Training Framework for Classification of Major Depressive Disorder using Non-IID Three Channel Electroencephalogram," Procedia Comput. Sci., vol. 235, pp. 56–69, 2024, doi: 10.1016/j.procs.2024.04.006.
[13] N. A. Abdelkadir, A. Ratnam, Z. Talat, and S. Chancellor, "FedMental: Evaluating Federated Learning for Mental Health Detection from Social Media Data," in Proc. ACL, 2026, pp. 40929–40944, doi: 10.18653/v1/2026.acl-long.1899.
[14] R. Kanauzia et al., "A comprehensive survey on federated learning for privacy preservation in digital healthcare applications," Knowl. Inf. Syst., vol. 68, art. no. 55, 2026, doi: 10.1007/s10115-025-02673-2.
[15] "DP-CARE: A differentially private classifier for mental health analysis in social media posts," Frontiers in Digital Health, 2025. [Study reports F1 78.08% and recall 88.67% on Dreaddit stress detection at ε≈3.]