

Comparative Analysis of Machine Learning Algorithms for Anxiety and Depression Risk Prediction

Sonal Shahi¹, Pradeep Yadav²

^{1,2} Department of Computer Science and Engineering, ITM Gwalior, Madhya Pradesh, India

Abstract: *The risk prediction of anxiety and depression can be considered one of the spheres of application of machine learning as the measurement of symptoms of anxiety and depression usually relies on self-reporting, interviews, or chance encounters. In this paper, comparative and literature review of classical machine learning algorithms and deep learning methods will be conducted to evaluate the use of machine learning approaches in risk prediction of anxiety and depression. These approaches include logistic regression, support vector machines, decision trees, random forest, gradient boosting, neural networks, convolutional neural networks, recurrent neural networks, transformers, and multimodal approaches. Unlike the traditional approach, which implies that the accuracy estimates obtained in the literature can be compared, in this review, several other aspects will be explored. The findings concerning risk prediction using electronic medical record information, social media textual data, speech information, and smartphone sensing information as well as other behavioral data indicate that tree-based algorithms and deep learning algorithms can be considered relevant for certain settings while having different performances in different studies depending on different factors. It means that the paper claims that algorithms should be chosen based on their applicability to the specific problem of risk prediction and deployment environment rather than just their accuracies. The framework for practical evaluation of the algorithms is developed, which takes into account such criteria as discrimination, calibration, robustness, fairness, interpretability, privacy, and clinical relevance.*

Keywords: anxiety prediction, depression prediction, machine learning, deep learning, explainable AI, digital phenotyping

1. Introduction

Anxiety and depression are some of the most common problems in the field of mental health that may impact involvement in education, employment, social interactions, sleep, and life in general. It may be difficult to implement preventive measures because risk factors may change over time between medical consultations. The availability of information in EHRs, questionnaires, language, social networks, and smartphones provides potential for developing algorithms that would detect signs which should lead to additional evaluation. Nevertheless, a signal by itself does not make a measure clinically valuable.

ML offers a number of techniques and not one technique. Traditional methods include logistic regression, support vector machine (SVM), decision tree, random forest (RF), and gradient boosting. These can be quite effective for structured data and fairly interpretable. On the other hand, neural techniques can extract more sophisticated features of text, speech, images, and longitudinal datasets, but they come with an increased need for data and computing resources for validation and explanation. A scoping review found that there were more than 300 research papers using ML/big-data techniques on mental health problems [1].

In this paper, we examine some common algorithms that have been used for risk prediction of anxiety and depression by means of a literature-based analytical approach. The key issue here is not about the highest accuracy of any algorithm, but rather the situation in which a certain class of models should be considered appropriate. The added value of our work lies in this comparative analysis.

2. Related Work and Evidence Base

Shatte et al. conducted an examination of the use of ML and big data in the field of mental health and identified 300 relevant articles for their systematic review based on a variety of uses, algorithms, and data types [1]. This demonstrates an issue which is common to systematic reviews: variations among studies in terms of targeted condition, type of data used, sample sizes, and methods of assessment make direct algorithm comparisons impossible. Recent systematic evidence further supports this claim. A review of text, audio, and video approaches for anxiety and post-traumatic stress from 2024 analyzed 128 articles [2].

As for depression, a 2023 systematic review on the use of AI to automatically diagnose depression by analyzing speech, text, and facial expression has been found, which has revealed that there is a significant amount of heterogeneous evidence in the field [3]. Moreover, a 2023 systematic review and meta-analysis of the application of wearables in depression have highlighted the importance of passive sensing technology despite certain design and generalizability issues [4].

The clinical context captured by EHRs includes diagnoses, procedures, medications, encounters, and clinical notes. Predictive models of depression and anxiety were created using EHR data by Nemesure et al., which showed that the use of clinical variables is feasible for prediction purposes [5]. This study also shows the importance of defining outcomes and the clinical setting due to the fact that the model built in one particular health system might reflect the documentation process.

HCET made use of hierarchical clinical embeddings and topic modeling in order to encode the heterogeneous EHR data. HCET showed an improvement in the precision-recall area under the curve metric of up to 0.07 against baseline models and it was found that multi-modal attention was important for performance and interpretability [6]. An EHR model based on transformers has recently been shown to produce a PRAUC score of 0.76 compared to 0.70 by the best baseline model tested in that experiment [7].

Data from text and behavior can be used to gain insights beyond the clinical domain. De Choudhury et al. used behavioral and linguistic characteristics from Twitter in order to assess depression risk prior to self-reporting of the condition [8]. Coppersmith et al. showed that it was possible to quantify the presence of mental health cues in Twitter language [9]. Eichstaedt et al. associated language on Facebook with depression as measured by medical records [10]. Reece and Danforth analyzed images from Instagram for predictive indicators of depression [11].

Digital phenotyping by means of smartphones is a method of extending digital phenotyping by employing passive sensing technology. The StudentLife study conducted by Wang et al. involved 48 students from a university over 10 weeks and correlated the behavior metrics derived from smartphones with stress, sleep, mood, socialization, and academic performance [12]. Torous et al. have provided an account of the technical framework required for conducting large-scale research on psychiatry using smartphones [13].

Anxiety prediction models have utilized questionnaire data, behaviors, text, speech, and video. According to the systematic review in 2024 by Kachkach et al., there have been 128 publications related to the prediction of anxiety and posttraumatic stress disorder based on text, audio, and video modalities, with 87 papers focusing on anxiety [2]. The review found that the promise of accurate predictions was there, but the methodological quality and reporting were not uniform. A result of high accuracy from a small dataset cannot be generalized as evidence of similar results in another group.

3. Methodology

This study employs a structured narrative approach rather than conducting an experiment anew. Peer-reviewed literature and review articles have been sourced through indexed publisher databases and bibliography, with special attention given to studies that directly assess ML in diagnosing anxiety disorders or depression, or a task similar to these mental health prediction tasks. Studies included in this review are mostly systematic reviews, empirical studies with data and evaluation criteria, and methods for reporting prediction models [1]–[7], [14], [15].

The comparison pulls out five characteristics: mode of data, type of algorithm, predictand, evaluation method, and described strengths/weaknesses. The numbers themselves are cited only when explicitly stated by the source itself. The performance numbers are shown as particular observations for that study rather than a consolidated benchmark. This is critical because the differences between studies in

prevalence, labeling, sample, features, prediction time frame, and validation can significantly affect performance.

These aspects include discrimination, calibration, sensitivity, specificity, precision, F1 score, ROC-AUC, and PR-AUC. When outcomes are unbalanced, PR-AUC and outcome-specific metrics may be more relevant than accuracy metrics. The TRIPOD+AI checklist gives updated reporting criteria for prediction models involving regression and artificial intelligence techniques [14], while the PROBAST+AI tool offers a structure for assessing the quality, risk of bias, and applicability [15].

The suggested workflow consists of data definition, followed by data preprocessing, leakage prevention, building baseline models, candidate models, separated validation, calibration, subgroup analysis, model explainability, and finally prospective evaluation. Patient level separation needs to be considered when there are multiple observations per patient. If there is any prospective application of the model, temporal splitting is preferred. External validation needs to be done before making generalization claims.

4. Comparative Analysis of Algorithms

1) Classical Machine Learning

Logistic regression is an excellent baseline technique for risk prediction with structures because its coefficients directly define the feature relationships according to model assumptions. The use of regularization techniques can prevent overfitting in cases where there are too many predictors. SVMs can create nonlinear classification boundaries using kernels and have been applied for text and medical classification tasks. Decision trees produce intuitive rules but can be unstable without pruning. Random forest classifiers combine decision trees into an ensemble and learn non-linear dependencies with little need for feature engineering. Gradient boosting techniques refine weak classifiers step-by-step and excel on tabular datasets. The weaknesses of such methods involve sensitivity to pre-processing, class imbalance, hyper-parameter configuration, and lack of transparency as compared to simpler linear models. When it comes to screening for mental health conditions, such classifiers have their appeal.

2) Deep Learning

CNNs have the ability to learn local patterns from either sequential data, images, or spectrogram-like data. RNNs, such as LSTM and bidirectional LSTMs, are meant to capture sequential relationships. Transformers use an attention mechanism to capture long-range dependencies and have been widely used on EHR sequences and language.

One of the benefits of deep learning is representation learning, which allows the model to learn features that would otherwise need manual design. One of the disadvantages of deep learning is that having a lot of capacity could actually increase overfitting and the difficulty in finding out what errors the algorithm is making. HCET and Transformer-based approaches to EHRs provide a way of incorporating temporal and heterogeneous data into depression prediction [6], [7].

Table I: Comparative Characteristics of Major Algorithm Families

Algorithm	Typical data	Complexity	Interpretability	Data need	Key consideration
Logistic Regression	Structured/ tabular	Low	High	Small-medium	Strong baseline; linear assumptions
SVM	Text/ tabular	Medium	Medium	Small-medium	Kernel/scaling sensitive
Random Forest	Tabular/ mixed	Medium	Medium	Medium	Robust; less transparent
Gradient Boosting	Tabular	Medium	Medium	Medium	Often strong on structured data
CNN	Speech/ image/ sequence	High	Low-medium	Medium-large	Local representation learning
LSTM/BiLSTM	Longitudinal sequence	High	Medium	Medium-large	Temporal dependencies
Transformer	Text/EHR sequence	Very high	Medium	Large	High capacity; data hungry
Multimodal DL	Mixed modalities	Very high	Low-medium	Large	Richer but complex fusion

5. Literature-Reported Performance

The published results highlight the need to interpret performance in context. Table II shows selected empirical

results. The numbers are not from a side-by-side comparison; they come from different populations, datasets, names, and validation processes.

Table II: Selected Literature-Reported Evidence

Source	Data	Target	Approach	Reported evidence
Nemesure et al. [5]	EHR	Depression + anxiety	ML	Study-specific predictive results
Speier et al. [6]	EHR + notes	Depression	HCET	PRAUC improvement up to 0.07 vs. baselines
Yang et al. [7]	Multimodal EHR	Depression	Transformer	PRAUC 0.76 vs. 0.70 baseline
Choudhury et al. [8]	Twitter	Depression	Statistical classifier	Signals before reported onset
Eichstaedt et al. [10]	Facebook language	Depression	ML/NLP	Language associated with later EHR diagnosis
Kachkach et al. [2]	Text/audio/video	Anxiety/ PTSD	Systematic review	128 studies; heterogeneous metrics
Abd-Alrazaq et al. [4]	Wearables	Depression	Meta-analysis	Performance varied across studies

One of the important observations for analysis is that rich modalities do not necessarily result in better evidence. This is shown by the HCET improvement of PRAUC, which is achieved through a specific way of representing EHR data and prediction [6], while the transformer method has been tested in a different environment [7]. This has been confirmed by the 2024 review on anxiety, which noted that lack of consistency in reporting made comparison difficult [2]. Algorithm selection should therefore follow the prediction objective and data-generating process.

6. Validation, Fairness, and Explainability

1) Validation and Generalizability

Prediction algorithms could show impressive results in internal validation but perform poorly outside of their environment. A review of the literature on clinical prediction algorithms for psychiatric diseases revealed that 94.5% of all developed models were assessed as being at high risk of bias, and only around one-fifth were validated externally [16]. The results imply that algorithmic complexity cannot make up for poor study design.

Some possible problems include leakage between the training and test data, repeating observations of the same individual in both samples, tuning of the algorithm on the test set, unrepresentative sampling, variables measured after the outcome, and inconsistent definition of the label. In case of anxiety/depression, the label could be defined through questionnaires, clinicians' diagnosis, ICD-codes in EHRs, or a surrogate variable.

2) Fairness and Subgroup Robustness

Models to predict mental health can reflect disparities since they are trained using data that is not representative of certain demographic, linguistic, cultural, and socio-economic groups. The same model may have a high AUC but low performance in a particular group. Therefore,

evaluations of such models should show sensitivity, specificity, and false positive/false negative behaviors specific to certain groups if the sample size allows.

3) Explainability

Explainable AI could make audits possible and make the results easier to understand by humans. For example, methods like SHAP can provide an explanation of the influence that each variable has on an individual prediction, while attention can help identify key variables in sequence-based models. Explanations cannot be used as evidence for causation. The feature can be relevant due to its correlation with some social or medical characteristic but not due to causation.

7. Privacy, Ethics, and Deployment

Such data are very sensitive. The collection of data should be relevant to the purpose stated, consent should be sought appropriately, identifiers should be protected, access restricted, and guidelines on how long data should be stored and when deleted be stated. Care must be taken during social media and mobile phone analysis since people might not expect normal activities to turn into mental health risk scores [13].

The best deployment model for a risk assessment algorithm would be human-centred screening. Predictions should raise an opportunity for proper analysis instead of tagging individuals based on predictions. People should know what data is collected, what predictions the model can make, and what will happen if a prediction is rated at high risk. The system should also define a pathway for correcting inaccurate data or predictions.

Utility of the operational use must be considered beyond prediction alone. Useful measures would include calibration, alert load, follow-up time, subgroup utility, provider

acceptability, and if the tool enhances the targeted process without creating any undue worry or stigma.

8. Discussion

The above comparisons suggest that there is no universal best ML algorithm for risk prediction of anxiety and depression. Classical approaches are still useful as they create clear benchmarks and may work well with structured data. The ensemble trees are suitable choices for heterogeneous tabular features, whereas deep learning algorithms become more preferable when the input contains complicated structures like sequences, languages, speeches, images, or multimodal data. Transformers are good options for high-dimensional longitudinal data, but their excess capacity needs justification due to the amount and quality of the data.

The main methodological implication is that accuracy alone is not enough to choose among models. In screening, failing to detect a patient at high risk and creating false positives have different implications. When probability predictions are presented, calibration becomes necessary. AUPRC might provide meaningful information when the outcome is unbalanced. Validation studies check how transportable the model is, while subgroup analysis checks for any potential bias in average performance. Explainability and privacy serve different purposes.

This point is reinforced by recent methodological guidelines. TRIPOD+AI emphasizes the need for clear reporting on the development and assessment of prediction models [14]. PROBAST+AI extends risk-of-bias and applicability assessment to AI prediction models [15]. A systematic review on the longitudinal course of depression and anxiety in 2026 found only nine eligible studies and highlighted a risk of overfitting due to the addition of too many predictors [17].

1) Practical Decision Matrix

Table III: Deployment-Oriented Model Selection

Setting	Candidate models	Primary reason	Critical test
Small structured dataset	Logistic regression/ RF	Interpretability + robustness	External validation
Moderate tabular EHR	RF / boosting	Nonlinear interactions	Calibration + subgroup tests
Longitudinal EHR	LSTM/ Transformer	Temporal representation	Temporal + external validation
Text-based screening	SVM/ Transformer	Language representation	Cross-language validation
Speech-based screening	CNN/ DL	Acoustic representation	Speaker-independent testing
Multimodal research	Fusion DL	Complementary signals	Ablation + modality robustness

9. Limitations and Future Research

This is a literature-based comparative study and is not a prospective clinical validation study. Published studies vary

in terms of population, labels used, dataset, definition of features, and evaluation process. Thus, the performance numbers should not be seen as a common leaderboard. Some recent evidence could also change with new datasets and changing reporting standards.

Future work should include benchmarking using labeled datasets that have demographic representation; doing external and prospective validation; reporting sensitivity, specificity, calibration, and AUPRC consistently; and uncertainty estimation. Multimodal models should be validated through ablation experiments to check whether each modality helps at all. Privacy-preserving learning, federated learning, and on-device inference may minimize the risk of data leakage but bring additional challenges for evaluation.

The other promising avenue is longitudinal prediction. Rather than just asking whether an individual meets a threshold right now, the models will predict their trajectory over time without suggesting that correlation implies causality. The human factors science will be required to understand how the risk signals provided by algorithms are being interpreted and used by professionals.

10. Conclusion

Machine learning provides many useful methods for detecting patterns related to anxiety and depression, but selection of algorithms should be based on the data modality, the goal of the prediction, number of patients in the study, and the setting where the model will be used. Logistic regression, SVM, random forests, and boosting are valuable benchmarks that may prove useful with structured data. CNNs, recurrent neural networks, and transformers have more expressive power for complex and sequential data types. However, there is much heterogeneity and consistent issues such as bias, overfitting, lack of external validation, and inconsistent reporting in the literature.

In this case, it is not enough just to get high accuracy when developing an early-risk system. Discrimination, calibration, robustness, fairness, interpretability, privacy, and clinical utility need to be assessed in concert. The correct use of machine learning in this case is to facilitate human-centred early detection of risk, but not replacement of professional evaluations.

References

- [1] A. B. R. Shatte, D. M. Hutchinson, and S. J. Teague, "Machine learning in mental health: A scoping review of methods and applications," *Psychological Medicine*, vol. 49, no. 9, pp. 1426–1448, 2019, doi: 10.1017/S0033291719000151.
- [2] M. Kachkach et al., "Use of machine learning algorithms based on text, audio, and video data in the prediction of anxiety and posttraumatic stress in general and clinical populations: A systematic review," *Biological Psychiatry*, vol. 96, no. 7, pp. 519–531, 2024, doi: 10.1016/j.biopsych.2024.06.002.

- [3] "A systematic review on automated clinical depression diagnosis," npj Mental Health Research, 2023, doi: 10.1038/s44184-023-00040-z.
- [4] A. Abd-Alrazaq et al., "Systematic review and meta-analysis of performance of wearable artificial intelligence in detecting and predicting depression," npj Digital Medicine, vol. 6, art. no. 84, 2023, doi: 10.1038/s41746-023-00828-5.
- [5] 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," Scientific Reports, vol. 11, art. no. 1980, 2021, doi: 10.1038/s41598-021-81368-4.
- [6] W. Speier et al., "HCET: Hierarchical Clinical Embedding with Topic Modeling on Electronic Health Record for Predicting Depression," 2021.
- [7] H. Yang et al., "Bidirectional Representation Learning from Transformers Using Multimodal Electronic Health Record Data to Predict Depression," IEEE Journal of Biomedical and Health Informatics, 2021, doi: 10.1109/JBHI.2021.3063721.
- [8] 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.
- [9] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in Twitter," in Proc. Workshop on Computational Linguistics and Clinical Psychology, 2014, pp. 51–60, doi: 10.3115/v1/W14-3207.
- [10] 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.
- [11] A. G. Reece and C. M. Danforth, "Instagram photos reveal predictive markers of depression," EPJ Data Science, vol. 6, art. no. 15, 2017, doi: 10.1140/epjds/s13688-017-0110-z.
- [12] R. Wang et al., "StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones," in Proc. ACM Int. Joint Conf. Pervasive and Ubiquitous Computing, 2014, pp. 3–14, doi: 10.1145/2632048.2632054.
- [13] J. Torous, M. V. Kiang, J. Lorme, and J.-P. Onnela, "New tools for new research in psychiatry: A scalable and customizable platform to empower data driven smartphone research," JMIR Mental Health, vol. 3, no. 2, e16, 2016, doi: 10.2196/mental.5165.
- [14] J. F. Cohen, P. M. M. Bossuyt, et al., "TRIPOD+AI: An updated reporting guideline for clinical prediction models," BMJ, vol. 385, q824, 2024, doi: 10.1136/bmj.q824.
- [15] K. G. M. Moons et al., "PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods," BMJ, vol. 388, e082505, 2025, doi: 10.1136/bmj-2024-082505.
- [16] "Clinical prediction models in psychiatry: A systematic review of two decades of progress and challenges," Molecular Psychiatry, 2022, doi: 10.1038/s41380-022-01528-4.
- [17] "Prediction models for longitudinal trajectories of depression and anxiety: A systematic review," Journal of Affective Disorders, vol. 401, art. no. 121255, 2026, doi: 10.1016/j.jad.2026.121255.
- [18] Y. Liang, X. Zheng, and D. D. Zeng, "A survey on big data-driven digital phenotyping of mental health," Information Fusion, vol. 52, pp. 290–307, 2019, doi: 10.1016/j.inffus.2019.04.001.