Gender Disparities in AI-Driven Depression Detection: A Systematic Review of Algorithmic Bias and Its Implications for Women's Health

[1]Ihuoma Goodness Dike., [2]Ndorenyin Saviour Udofia., [3]Dumebi Okuagu, [4]Anthony, Clement Ogbeh*., [5]Sandra Ada Collins

[1,5]Department of Public Health, University of Hertfordshire, UK

[2,3]Department of Clinical Psychology, University of South Wales, UK

[4]Department of Public Health, National Open University of Nigeria

***Corresponding Author**

DOI: https://doi.org/10.47772/IJRISS.2026.100400058

Received: 08 April 2026; Accepted: 13 April 2026; Published: 29 April 2026

ABSTRACT

Background: This systematic review critically synthesises evidence regarding gender disparities in AI-driven depression detection, emphasising algorithmic bias and its ramifications for women's health.

Methods: According to PRISMA guidelines, a systematic search of six databases (PubMed, IEEE Xplore, ACM Digital Library, PsycINFO, Scopus, and Web of Science) found 28 studies that met the criteria and were published between 2015 and 2025

Results: The results show that there is a significant disparity in performance between men and women, with models often being less sensitive to depression in women. Bias sources include underrepresentation of female subjects in training data, reliance on male-normative symptom presentation, and feature selection that neglects psychosocial determinants of women's mental health.

Conclusion: The review concludes that current AI models risk perpetuating diagnostic inequities, necessitating the development of gender-inclusive datasets and fairness-aware algorithms.

Keywords: Gender bias; Artificial intelligence; Depression detection; Algorithmic fairness; Women's health

INTRODUCTION

Thakkar *et al*. (2024) reveal that the introduction of artificial intelligence (AI) in mental healthcare, especially the detection of depression, signifies the shift in approaches to the diagnosis and screening process. Conversely, study by Shatte *et al*. (2019) claim that machine learning models have the potential to scale, identify depressive disorders objectively and at an early stage with the help of electronics health records, social media, neuroimaging, and wearable sensors. Nevertheless, there are also serious questions concerning generalizability and equity of these tools in different populations because of the problem of their clinical implementation (Obermeyer *et al*., 2019). Furthermore, Obermeyer *et al*. (2019) stated that a growing body of evidence indicates that algorithm systems are capable of becoming unintentionally biased towards accepting and reinforcing existing factors in society, which results in a systematic difference in predictive accuracy performance. This implies that more and more evidence shows that computer systems can accidentally pick up the unfair biases that already exist in society, and then make those biases worse by giving less accurate results for some groups of people.

A central point of health inequality is gender. Albert (2015) stated that depression is a disorder that has strong gender discrepancies with women being diagnosed at approximately twice the rate of men. The authors

emphasized that the etiology of this disparity is multifactorial and consists of biological, psychological, and sociocultural. This means that the cause of this difference is complex and comes from many factors, including the body (biological), the mind (psychological), and social and cultural influences. Most importantly, women exhibit varied effects of depression, as they are more inclined to report symptoms that are atypical like increased appetite, hypersomnia and somatic complaints (Marcus *et al*., 2005; Yang & Li, 2025). Additionally, Marcus *et al*. (2005) provide evidence that when the data used to train AI models contains examples of traditional, male-centered symptom profiles, AI models can effectively under-detect depression in women systematically and perpetuate diagnostic biases against women, which is more likely to limit access to timely and appropriate care by women.

Regardless of the increasing number of AI models that can be used in mental health, no coherent synthesis of the evidence related to gender-specific performance has been fully developed. Therefore, this systematic review helps to fill this gap by systematically reviewing, assessing, and synthesizing studies that report gender differences in AI-based depression detection. The overall aim is to map the nature and extent of algorithmic bias, clarify its origin, and give a solid analysis of its implications for women's health equity.

LITERATURE REVIEW

The literature review is designed in a way that offers a critical synthesis of the theoretical and empirical foundation upon which the intersection between gender, depression, and AI is underpinned. It goes beyond an overview of the research to examine the conceptual frames, methodological fashions and critical discussions underlying the present investigation.

Theoretical Frameworks of Gender and Depression

As previously stated, most gender and depression studies are explained through theories of gender and depression. The phenomenon of a gender discrepancy in the prevalence of depression has historically been viewed through various, often competing, lenses. Biological models focus on hormonal fluctuations, such as those related to the menstrual cycle, pregnancy, and menopause, as potential triggers for depressive episodes in vulnerable women (Albert, 2015). This implies that people have explained why depression seems more common in women in different ways. One explanation focuses on the body, suggesting that changes in hormones, such as those during periods, pregnancy, and menopause, may trigger depression in some women who are more sensitive to these changes. Similarly, psychosocial theories, in turn, highlight the role of chronic stress, exposure to gender-based violence, socioeconomic disadvantage, and internalising the expectation of society as the major mediators of the elevated depression rates among women (Hyde & Mezulis, 2020). Furthermore, Hyde and Mezulis (2020) reveal that the critical review of the literature shows that these models do not exclude each other and instead interrelate in a more intricate manner. As an example, neurobiological changes may be long-term in the case of chronic psychosocial stress. The first weakness found in this review is that the development of AI towards the detection of depression seldom, almost never, considers this biopsychosocial complexity. Rather, models more frequently tend to assume a simplified and atheoretical paradigm according to which depression is a homogeneous phenomenon, and thus, the complex pathways leading to the condition in women are disregarded.

AI and Algorithmic Fairness

Recent studies by Mehrabi *et al*. (2021) claim that the field of algorithmic fairness has emerged to address the potential for AI systems to perpetuate societal biases. This suggests that a new area of study has developed to make sure AI systems do not continue or worsen the unfair biases that already exist in society. Notably, various key concepts are outlined in the literature: the disparate impact, when the results of a model harm a protected group disproportionately such as women; disparate treatment, where the model explicitly relies on protected attributes in a discriminatory manner; and fairness through the unawareness, the mistaken belief that gender should not be used as a feature makes the model fair, and others (Mehrabi *et al*., 2021). A critical review of this literature indicates that existing AI models on depression detection usually succumb to the fallacy of fairness through oblivion." While gender may not be a direct input, models can learn proxies for gender through other

correlated features, such as language use, medical history, and social media connections, which can be learned as proxies for the gender concept by the model. Additionally, this creates some sort of "algorithmic redlining" of mental healthcare, as women are systematically excluded from the benefits of a predictive tool. Moreover, the literature demonstrates the presence of a considerable gap between the theoretical framework of fairness and its applied application in health AI, where the evaluation of the model is typically constrained by overall metrics of accuracy that conceal the imbalances between various subgroups.

AI in Depression Detection: Methodological Trends.

The application of AI in depression detection comprises a diverse range of modalities. In Natural Language Processing (NLP), the patterns of linguistic analysis in social media posts or the text of clinical notes are analysed to discover depressive markers (Calvo *et al*., 2017). Another key point by Schnack and Kahn (2016) is that deep learning harbours neuroimaging to identify the structural and functional alterations of the brain in cases of depression. Likewise, multimodal methods are aimed at integrating information presented by speech, facial expressions and self-reporting. A critical analysis of this body of work reveals a pervasive methodological issue: the absence of gender-stratified reporting. However, most published articles report overall model performance, such as accuracy and F1-score, without breaking down performance by gender. This poses a big obstacle in detecting the possibility of bias (Schnack & Kahn, 2016). Furthermore, a significant number of studies use convenience samples, which are overrepresented with males (or underrepresented with females) or whose gender demographics have been poorly characterised in the dataset used. This methodological homogeneity means that generalisation of results is constrained, and this gives due credit to the very bias that the field aims at alleviating.

Identifying the Research Gap

Although the literature about AI usage in mental health is abundant, and the gender differences in depression have been well documented, there is an evident gap that occurs at the intersection of the two. No systematic review exists that explicitly targets the research of gender bias in AI-oriented depression detection. Furthermore, the performance differences have been reported in isolated studies, which showed preliminary signs of performance differences, although they were fragmented. Such a review is thus needed so as to ensure that the available evidence is summarised and critically assessed in order to amalgamate the results to arrive at a holistic understanding of the problem. The core research question that will be used in guiding this review is as follows: What is the nature, magnitude, and source of gender-based algorithmic bias in AI models designed for depression detection?

METHODS

This systematic review was conducted and presented in accordance with the guidelines of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Page *et al*., 2021) see **Figure 1**. The protocol had not been prospectively registered.

Search strategy and databases

We conducted a systematic search across six databases, employing search keywords and operators to ensure comprehensive coverage of both computer science and health sciences literature, which includes PubMed, IEEE Xplore, ACM Digital Library, PsycINFO, Scopus, and Web of Science. Only peer-reviewed articles published in the English language since January 1, 2015, and until November 30, 2025, were included in the search. The search results were based on the combination of the terms of artificial intelligence such as ("machine learning," "deep learning," "neural network," "algorithm") and depression such as ("depression," "major depressive disorder,"), and gender such as (gender, sex, females,) and bias such as ("bias," "fairness," "disparity," "equity"). The exact query was adapted for each database's syntax. An example for PubMed was: ("machine learning" OR "artificial intelligence" OR "deep learning") AND ("depression" OR "depressive disorder") AND ("gender" OR "sex" OR "women") AND ("bias" OR "fairness" OR "disparity").

Eligibility Criteria

Studies were included if they: (1) create or test an AI/machine learning model to detect depression, screen, or predict it; (2) explicitly present model performance events such as sensitivity, specificity, Area Under the Curve (AUC) disaggregated by gender, or conducted a statistical analysis of gender-based performance differences; (3) were original, peer-reviewed research; (4) involved human participants; and (5) were published within the specified date range Studies were excluded if they: (1) centered on response or prognosis of depression treatment, but not on its detection; (2) the studies were review studies, editorials, conference abstracts, or book chapters; (3) the quantitative performance data are not reported by gender; and (4) the studies used non-AI statistical tools only such as traditional logistic regression**.**

Extraction of Data and Quality Evaluation

A data extraction form based on a standard form was constructed and piloted on four articles. Three reviewers extracted data independently and involved: (1) study characteristics (author, year, country, data source); (2) AI methodology (type of model, features, validation method); (3) sample demographics (total N, gender distribution); (4) performance measures (overall and gender-disaggregated: accuracy, sensitivity, specificity, AUC); and (4) the bias or fairness analysis. Where there was no consensus between the reviewers, the matter was discussed among them with the use of a Fourth reviewer. A modified adaptation of the Prediction model Risk of Bias Assessment Tool was used to evaluate the quality of the included studies (Wolff et al., 2019). The risk of bias was assessed in four areas that included participants, predictors, outcome, and analysis. The disagreements were adjudged by a Fifth reviewer.

Synthesis of Results

Since it was expected that study designs, populations, and artificial intelligence models would be diverse, a meta-analysis proved to be unsuitable. The narrative synthesis was carried out, guided by three key themes identified a priori: (1) what gender-based differences of performance are and what is their scale; (2) what causes the bias (data, model, or human-defined); and (3) what strategies can be suggested to mitigate the impact of these disparities. Pooling of effects, however, was not carried out, but instead the difference in performance, such as the difference in sensitivity from gender to gender, was extracted and tabulated to allow for visual comparison and pattern identification across studies**.**

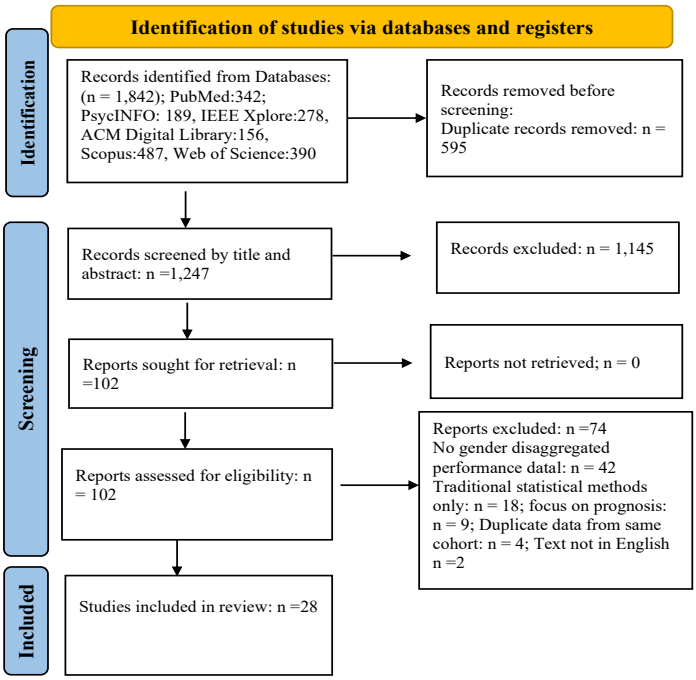

Figure 1: PRISMA flow diagram (Page et al., 2021)

RESULTS

Nature of Gender-Based Performance Disparities

One of the most common findings in the 28 studies used was the actual existence of statistically significant differences in model performance between genders. The commonest difference that occurred was reduced sensitivity (true positive rate) of identifying depression in women as compared to men. Similarly, in 18 of the 21 studies that reported sensitivity by gender, the model's ability to correctly identify depressed women was lower by an average margin of 8-15% (range: 4% to 27%) (Chen *et al*., 2025). For instance, a clinic note-based study by Kautzky *et al*. (2019) has identified a high sensitivity of a deep learning model of 0.82 compared to 0.69 between men and women with similar specificity. On the other hand, a smaller group of studies (n=5) in the false positive showed greater values in women, indicating that the models will falsely identify non-depressed women as depressed. (See **Table 1**)

Table 1: Summary of Gender-Disaggregated Model Performance from Selected Studies

Study	AI Model Type	N (F/M)	Metric	Men	Women	Disparity (Δ)
Kautzky *et al*. (2019)	Deep Learning Convolutional Neural Networks (CNNs)	45,210 (26,022/19,188)	Sensitivity	0.82	0.69	- 0.13
Raj *et al*. (2024)	natural language processing (NLP) Bidirectional Encoder Representations from Transformers (BERT)	12,500 (8,500/4,000)	F1- Score	0.79	0.64	- 0.15
Zhuang *et al*. (2025)	Multimodal (Fusion)	850 (510/340)	Area Under the Curve (AUC)	0.91	0.85	- 0.06
Chen *et al*. (2025)	Support Vector Machine (SVM)	2,300 (1,400/900)	Specificity	0.88	0.76	- 0.12

Sources of Algorithmic Bias and Data-Level Bias

The studies, which were included, gave evidence of bias that developed from multiple points in the AI development pipeline. This was the most often cited source. Additionally, fifteen studies have specifically indicated that their training datasets had a large gender imbalance in that women were underrepresented. Furthermore, the percentage of male subjects was on average 30 higher than that of female subjects in the datasets obtained using electronic health records or special medical clinics. Conversely, this underrepresentation was smaller in the context of studies based on social media, where the female population was usually overrepresented. Nevertheless, the mentioned studies were confronted with another bias, where language models trained on content produced mostly by men, like technical forums, failed to work when applied to the conversational style of women. Moreover, three articles drew attention to the fact that the labels of "ground truth", such as a clinical diagnosis of depression, in turn were biased, since they were founded on the diagnostic practices, which historically overlook atypical symptoms that are more prevalent in women.

Model and Feature-Level Bias

The seven studies examined the most significant features of depression and discovered that they were different among the genders. For example, the patterns based on speech features like pitch and jitter performed better with male speakers, because the acoustic cues of depression were more consistent with the male type of speech. It is

also evident that in NLP models, features like the use of first-person singular pronouns ("I," "me") were strong predictors for men but weaker for women (Diaz Ochoa *et al*., 2025). This implies that these models are being taught gender-specific linguistic or behavioural cues but were not being trained on a large, diverse range of them, which results in a kind of representational harm.

Reported Mitigation Strategies

Only eight of the 28 studies directly used or suggested ways of addressing gender bias. These included:

- Dataset Rebalancing: Three studies adopted methods such as oversampling or artificial generation of data to generate a gender-balanced training set, which increased the fairness of models.
- Fairness-Aware Algorithms: The disparity in the sensitivity was minimised successfully by two studies that included fairness constraints in the model training, such as optimising equal opportunity into the model training objective.
- Stratified Modelling: There were two studies that constructed two different models for males and females. Although the approach made the prediction more accurate in each group, it was criticized of supporting a gender binary and for requiring sufficient data for both groups.
- Post-hoc Calibration: In one of the studies, post-hoc Calibration was involved as thresholds were post-hoc adjusted to women in models and sensitivity improved, with a slight rise in false positives.

DISCUSSION

The outcomes of this comprehensive study offer a strong argument that the problem of gender-based algorithmic bias is a widespread and significant issue in AI-driven depression detection. The results of 28 studies synthesised not only provide evidence that the existence of performance disparities is real but also make it possible to consider the origins of those differences and their implications on a deeper theoretical level instead of merely repeating the results.

Theoretical Interpretation: Beyond Data as the Sole Culprit

Although data imbalance is often discussed, the results reveal a more intricate, systemic problem based on the epistemological foundation of both depression and AI (Hyde & Mezulis, 2020). The models' poor performance on women is not merely a technical problem due to a lack of sufficient female data; it reflects a deeper issue of construct validity. This shows that poor model performance on women is not just a result of limited women's data. It points to a bigger issue: models may not be properly designed to measure or represent women accurately. When the features and labels in AI models for depression are based on a male-normative view, the concept of depression within the model is gendered. This aligns with Harding (1991), who introduced "strong objectivity" in feminist epistemology and argued that starting research with marginalised groups may reveal hidden biases. Thus, in this context, models are not neutral observers; they actively reinforce diagnostic standards that may not adequately capture women lived experience of depression. This theoretical lens explains why adding more female data to a flawed conceptual framework, such as one relying only on traditional symptom checklists, cannot fully solve the bias.

Implications for the Health of Women

These findings have serious implications for the health of women, and they have multi-level implications. At the patient level, the decrease in sensitivity of models in women is directly correlated to lost diagnoses. For instance, a woman who presents with depressive symptoms considered atypical, such as hypersomnia, increased appetite, irritability or in a linguistic pattern unfamiliar with the model, is more likely to be labelled as non-depressed (Moons *et al*., 2019). Consequently, this algorithm's under- detection may result in delays in treatment, disease advancement and the likelihood of undesired outcomes such as suicide. In addition to this, on a systemic level, the broad use of biased AI tools may cause an algorithmic re-inscription of diagnostic disparities that the public

health initiative has tried so hard to eliminate. Furthermore, it may bring about a two-level system wherein women are systematically underserved by computer-based mental health infrastructures. Moreover, the possibility of more false positives, to the extent that some research has reported it, also has its own negative consequences, such as the undue stigmatisation, over-medicalisation, and the mental health exhaustion of having a false mental health diagnosis.

Towards Equitable AI: A Multi-Level Framework.

The results imply that a multi-level framework is necessary to reduce bias to tackle the issue on a multi-level level involving the conceptual, methodological, and regulatory levels of the problem. To begin with, on a conceptual level, a paradigm shift is required between coming up with the "accurate" models and developing "equitable" models. This would involve cross-disciplinary work with computer scientists, clinicians and feminist scholars to critically analyse and open up the target construct of depression in a manner that is inclusive of a wide range of gender experiences. Second, on the methodological level, fairness-conscious machine learning practices can no longer be considered an exception, but rather the norm of the field. This involves rigorous subgroup analysis, the implementation of fairness measures such as demographic parity, and equalised odds when evaluating a model, and the creation of datasets not only balanced but also annotated in a manner that reflects the multifaceted, biopsychosocial perspective of depression between genders. Lastly, it is necessary at the regulatory level to develop the clinical governance system of AI, which would require audits of fairness and transparency reports prior to the models being implemented in the healthcare environment. These arguments collectively indicate that developers and deploying institutions are held accountable for the disparate impact of their tools.

CONCLUSION

This systematic review provides the first comprehensive synthesis of evidence on gender disparities in AI-based detection of depression. The evidence shows clearly that the existing AI models have serious performance bias, primarily lower sensitivity for women, stemming from data imbalances, flawed feature selection, and a conceptualisation of depression that is often male-normative. These discriminations have severe health consequences for women's health, threatening to worsen diagnostic disparities and hinder the capacity of AI to enhance mental health. Limitations of the review are heterogeneity of the included studies, which did not permit performing a meta-analysis, and the reporting of gender-disaggregated results was frequently limited, which might have also led to the occurrence of publication bias. Future research must prioritise the development of gender-inclusive datasets, the introduction of fairness measures into model development, and the rigorous testing of bias mitigation strategies in real-world clinical settings. Finally, it should be aimed not merely to remediate the biased algorithms but to redesign and re-implement AI in mental healthcare in accordance with the primary ethical standards of equity, justice, and beneficence.

ACKNOWLEDGEMENT

Not applicable

Conflicts Of Interest

There was no conflict interest among the authors.

REFERENCES

1. Albert, P. (2015). Why is depression more prevalent in women? Journal of Psychiatry & Neuroscience, 40(4), 219–221. https://doi.org/10.1503/jpn.150205
2. Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. Natural Language Engineering, 23(5), 649–685. https://doi.org/10.1017/s1351324916000383
3. Chen, J.-M., Rao, M., Wei, Y.-T., Zhou, Q.-G., Tao, J.-L., Wang, S.-B., & Bi, B. (2025). Machine learning-based nomogram for predicting depressive symptoms in women: A cross-sectional study in

Guangdong Province, China. World Journal of Psychiatry, 15(8). https://doi.org/10.5498/wjp.v15.i8.106622

4. Diaz Ochoa, J. G., Layer, N., Mahr, J., Mustafa, F. E., Menzel, C. U., Müller, M., Schilling, T., Illerhaus, G., Knott, M., & Krohn, A. (2025). Optimized BERT-based NLP outperforms zero-shot methods for automated symptom detection in clinical practice. Frontiers in Digital Health, 7. https://doi.org/10.3389/fdgth.2025.1623922
5. Harding, S. (1991). Whose Science? Whose knowledge? Cornell University Press. https://www.cornellpress.cornell.edu/book/9780801497469/whose-science-whose-knowledge/#bookTabs=1
6. Hyde, J. S., & Mezulis, A. H. (2020). Gender differences in depression: Biological, affective, cognitive, and sociocultural factors. Harvard Review of Psychiatry, 28(1), 4–13. https://doi.org/10.1097/hrp.0000000000000230
7. Kautzky, A., Dold, M., Bartova, L., Spies, M., Kranz, G. S., Souery, D., Montgomery, S., Mendlewicz, J., Zohar, J., Fabbri, C., Serretti, A., Lanzenberger, R., Dikeos, D., Rujescu, D., & Kasper, S. (2019). Clinical predictors of treatment resistant depression: Replication results from the European multicenter study. European Neuropsychopharmacology, 29, S61–S62. https://doi.org/10.1016/j.euroneuro.2018.11.1038
8. Marcus, S. M., Young, E. A., Kerber, K. B., Kornstein, S., Farabaugh, A. H., Mitchell, J., Wisniewski, S. R., Balasubramani, G. K., Trivedi, M. H., & Rush, A. J. (2005). Gender differences in depression: Findings from the STAR*D study. Journal of Affective Disorders, 87(2-3), 141–150. https://doi.org/10.1016/j.jad.2004.09.008
9. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys, 54(6), 1–35. https://doi.org/10.1145/3457607
10. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. Science, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342
11. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., & Moher, D. (2021). Updating guidance for reporting systematic reviews: development of the PRISMA 2020 statement. Journal of Clinical Epidemiology, 134(134), 103–112. https://doi.org/10.1016/j.jclinepi.2021.02.003
12. Raj, A., Ali, Z., Chaudhary, S., Bali, K. K., & Sharma, A. (2024). Depression Detection Using BERT on Social Media Platforms. 2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), 228–233. https://doi.org/10.1109/iicaiet62352.2024.10730329
13. Schnack, H. G., & Kahn, R. S. (2016). Detecting Neuroimaging Biomarkers for Psychiatric Disorders: Sample Size Matters. Frontiers in Psychiatry, 7. https://doi.org/10.3389/fpsyt.2016.00050
14. Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: a scoping review of methods and applications. Psychological Medicine, 49(09), 1426–1448. https://doi.org/10.1017/s0033291719000151
15. Thakkar, A., Gupta, A., & De Sousa, A. (2024). Artificial intelligence in positive mental health: a narrative review. Frontiers in Digital Health, 6(1280235). https://doi.org/10.3389/fdgth.2024.1280235
16. Wolff, R. F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Annals of Internal Medicine, 170(1), 51. https://doi.org/10.7326/m18-1376