Digital Neuropsychiatry
eISSN: 3122-6671
https://dn.cultechpub.com/index.php/dn

From Clinic to Code: Developing Interpretable ML Models for Treatment Selection in Major Depressive Disorder

Mercado Protacio

Department of Psychiatry, University of Fatima, Valenzuela City, Metro Manila, Philippines

Abstract

Background: Major Depressive Disorder (MDD) is a heterogeneous condition with a wide range of treatment options, yet achieving remission remains a challenge due to the trial-and-error nature of treatment selection. While machine learning (ML) promises to personalize this process, "black-box" models often lack clinical trust and actionable insights, limiting their adoption.

Objective: This study aims to develop and validate an interpretable ML pipeline for predicting optimal first-line treatment selection between Selective Serotonin Reuptake Inhibitors (SSRIs) and Cognitive Behavioral Therapy (CBT) for patients with MDD.

Methods: We utilized a dataset of 1,250 patients from the [Anonymized] Neuropsychiatric Dataset, featuring comprehensive clinical, demographic, and digital phenotyping data. We trained and compared several ML models, including a black-box Gradient Boosting Machine (GBM) and an interpretable Explainable Boosting Machine (EBM). Model performance was assessed using accuracy, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC). Interpretability was achieved through global feature importance and local explanations for individual predictions.

Results: The GBM model achieved the highest performance (AUC = 0.87), with the EBM model performing comparably (AUC = 0.85). Crucially, the EBM provided transparent insights, identifying key predictors of treatment success, such as baseline anxiety severity, sleep disturbance patterns, cognitive performance scores, and actigraphy-derived physical activity levels. A novel, clinically-actionable visualization, the "Treatment Suitability Scorecard," is presented for individual patient guidance.

Conclusion: Interpretable ML models can achieve performance comparable to black-box models while providing crucial transparency for treatment selection in MDD. The proposed pipeline and visualization tools facilitate the transition of ML from a research tool to a clinically-deployable decision-support system, fostering trust and enabling personalized, evidence-based care in digital neuropsychiatry.

Keywords

Major Depressive Disorder, Interpretable Machine Learning, Treatment Selection, Explainable AI, Digital Phenotyping, Precision Psychiatry, Cognitive Behavioral Therapy, Antidepressants

1. Introduction

Major Depressive Disorder (MDD) is a leading cause of disability worldwide, characterized by its profound heterogeneity in symptomatology, underlying biology, and treatment response. The current standard of care for treatment selection is largely empirical, following evidence-based guidelines but ultimately relying on a trial-and-error approach. It is estimated that only approximately 30% of patients achieve remission with their first antidepressant trial, with subsequent trials yielding diminishing returns [1]. This protracted process prolongs patient suffering, increases the burden on healthcare systems, and contributes to treatment-resistant depression.

The field of computational psychiatry has emerged with the promise of leveraging data-driven methods to overcome this challenge. Machine learning (ML) models, capable of identifying complex, multivariate patterns in large datasets, offer a pathway to personalize treatment decisions. Early proof-of-concept studies have demonstrated that ML can predict general treatment outcomes with modest accuracy. However, a significant translational gap remains. The most powerful predictive models, such as deep neural networks and ensemble methods, often operate as "black boxes," providing little to no insight into *why* a particular prediction was made [2].

For clinicians to trust and effectively utilize an ML recommendation, they require more than a probability score. They need to understand the clinical rationale behind it. Was the recommendation driven by a specific symptom cluster? A

biomarker? A social determinant? Without this transparency, clinicians are rightfully hesitant to integrate these tools into their high-stakes decision-making process. This has spurred a critical movement towards *interpretable* or *explainable* machine learning in healthcare (Rudin, 2019).

This study addresses this gap by developing and validating an interpretable ML pipeline specifically for the task of initial treatment selection in MDD. We focus on the common first-line choice between pharmacotherapy (SSRIs) and psychotherapy (CBT). We posit that an interpretable model will achieve predictive performance comparable to a black-box model while providing clinically meaningful insights that can directly inform patient-physician dialogue and personalize treatment plans. By moving "from clinic to code," we aim to bridge the chasm between statistical prediction and clinical action, paving the way for the responsible implementation of AI in digital neuropsychiatry [3].

2. Theoretical Background and Literature Review

2.1 The Heterogeneity of MDD and the Imperative for Personalization

The failure of a one-size-fits-all approach in MDD treatment is rooted in the disorder's etiological and phenotypic diversity. The Research Domain Criteria (RDoC) framework explicitly acknowledges this, encouraging a multi-level understanding of mental disorders spanning from genomics to self-report. This heterogeneity means that patients who present with similar total scores on the Hamilton Depression Rating Scale (HAM-D) may have vastly different symptom profiles (e.g., anhedonia vs. anxiety/somatization), which are known to predict differential response to various treatments [4]. This phenotypic diversity is mirrored by heterogeneous neurobiological underpinnings. Neuroimaging studies have consistently failed to identify a single "depression circuit," instead revealing alterations across multiple brain networks including the default mode, salience, and cognitive control networks. The implication for treatment is profound: a patient with prominent anhedonia and blunted reward-system activity may constitute a neurobiological subtype that responds differently to a dopamine-focused intervention compared to a patient with overwhelming anxiety and hyperactive amygdala reactivity. This biological rationale provides a compelling foundation for moving beyond syndromal classification towards a data-driven, mechanistically-informed approach to treatment personalization [5].

2.2 Machine Learning in Psychiatry: From Prediction to Clinical Utility

Early ML applications in psychiatry focused primarily on diagnostic classification (e.g., distinguishing MDD from bipolar disorder) using neuroimaging data. More recently, the focus has shifted to predicting treatment outcomes. For instance, a landmark study by Chekroud et al. (2016) used a large dataset from the STAR*D trial to predict remission to citalopram, identifying features like employment status and sleep quality as predictors. However, such models often aggregate data across multiple treatment steps, lacking specificity for the initial selection between distinct modalities like medication and therapy. Furthermore, a critical review of the literature reveals a common limitation: a predominant focus on predicting response to a single treatment, most often pharmacotherapy. While valuable, this approach does not directly address the clinician's fundamental dilemma of choosing between different treatment pathways. A smaller but growing body of work has begun to tackle differential treatment prediction. For example, some studies have used EEG signatures to predict SSRI response over placebo, while others have explored linguistic features from clinical interviews as predictors of psychotherapy outcomes. Our study builds directly upon this nascent literature by explicitly modeling the comparative effectiveness of two first-line interventions with distinct mechanisms of action, thereby providing a more directly actionable tool for the point of care [6].

2.3 The Interpretability Imperative in Clinical ML

The demand for interpretability is not merely academic; it is a practical and ethical necessity. Rudin (2019) compellingly argues that black-box models are problematic for high-stakes decisions because they can be unstable, encode biases, and are unaccountable. In psychiatry, where the therapeutic alliance is paramount, a model that can explain its reasoning can serve as a collaborative tool rather than an opaque oracle [7]. Techniques like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) have been developed to provide post-hoc explanations, but inherently interpretable models (e.g., linear models, decision trees, and EBMs) are often preferable as their internal logic is transparent by design. The choice of an interpretable model like the Explainable Boosting Machine (EBM) is thus a deliberate one, grounded in the principle of "interpretability by design." Unlike post-hoc methods that approximate a black-box model's behavior, EBM's are Generalized Additive Models (GAMs) that learn individual shape functions for each feature, making the contribution of every variable directly visible and understandable. This allows a clinician to not only see which factors are important, but also the nature of their relationship with the outcome-for instance, whether the probability of CBT success increases linearly with anxiety severity or only beyond a specific threshold. This level of transparency is crucial for generating testable clinical hypotheses and fostering genuine trust.

2.4 Digital Phenotyping as a Source of Predictive Features

Digital phenotyping-the moment-by-time quantification of individual-level human phenotype using data from personal digital devices-offers a rich, objective source of data for ML models. Features derived from smartphone sensors (GPS, accelerometer), keystroke dynamics, and voice analysis can provide ecologically valid markers of sleep, mobility, social interaction, and cognitive motor integration, which are central to the symptomatology of depression. Integrating these "real-world" data streams with traditional clinical measures holds great promise for creating more robust and

personalized prediction models. For instance, GPS-derived location entropy (a measure of movement randomness and routine) has been shown to correlate with anhedonia and negative symptoms, while actigraphy-derived sleep efficiency provides an objective measure of insomnia severity that may be less susceptible to recall bias than self-report [8]. By incorporating these objective, continuous measures, our model moves beyond the snapshot provided by a clinical interview, capturing behavioral manifestations of depression as they unfold in the patient's natural environment. This enriches the feature space with variables that are both clinically relevant and computationally tractable, potentially capturing aspects of the illness that are not fully articulated by the patient or captured by standard rating scales.

3. Methodology

3.1 Study Design and Participant Selection

This study is a secondary analysis of data from the [Anonymized] Neuropsychiatric Dataset, a longitudinal, naturalistic study of mood and anxiety disorders. The analysis cohort consisted of 1,250 adult patients (age ≥ 18) with a primary diagnosis of MDD, confirmed by the Structured Clinical Interview for DSM-5 (SCID-5). All participants were treatment-naïve or had been off psychotropic medications for at least 4 weeks prior to baseline assessment. Participants were randomized to receive either a standardized SSRI (escitalopram) or a manualized, 16-week CBT protocol [9].

3.2 Measures and Feature Engineering

The primary outcome was treatment response, defined as a ≥50% reduction from baseline HAM-D score at week 8. Remission (HAM-D ≤ 7) was a secondary outcome.

A wide range of baseline features were extracted and engineered into five domains:

- **Demographics & Clinical History:** Age, gender, age of onset, number of previous episodes.
- **Symptomatology:** HAM-D total score and subscales (e.g., sleep, anxiety, weight), Beck Anxiety Inventory (BAI) score.
- **Cognitive Assessment:** Digit Symbol Substitution Test (DSST), Trail Making Test Part B (TMT-B).
- **Digital Phenotyping (2-week passive monitoring):**
  - **Actigraphy:** Sleep efficiency, total sleep time, daytime activity variance (derived from a wrist-worn accelerometer).
  - **Smartphone Use:** Number of outgoing calls, total screen-on time, location entropy (a measure of movement diversity).
- **Genetics:** Polygenic Risk Score for MDD (PRS-MDD).

3.3 Data Preprocessing and Feature Selection

Prior to model training, a rigorous data preprocessing pipeline was implemented. Missing data, which constituted less than 5% of the dataset, was handled using multivariate imputation by chained equations (MICE), under the assumption that data was missing at random. All continuous features were standardized (z-score normalization) to ensure comparability of coefficients and importance scores. To mitigate the risk of overfitting and enhance model generalizability, we employed a two-stage feature selection process. First, a univariate analysis (ANOVA F-test for continuous features, chi-square for categorical) was conducted to filter out features with no significant association with the treatment response outcome ($p > 0.10$). This was followed by a recursive feature elimination (RFE) procedure with cross-validation on the training set to identify the most parsimonious set of predictors that maintained optimal model performance. This process refined our initial feature set from 45 to 28 core variables for the final modelling [10].

3.4 Machine Learning Pipeline

The data was split into a training set (70%) and a held-out test set (30%). A nested cross-validation approach was used on the training set for hyperparameter tuning and model selection to prevent data leakage and overfitting. The following models were implemented:

- **Logistic Regression (LR):** A simple, interpretable baseline.
- **Random Forest (RF):** A robust ensemble method.
- **Gradient Boosting Machine (GBM):** A high-performance black-box model.
- **Explainable Boosting Machine (EBM):** A high-performance, inherently interpretable model based on generalized additive models.

3.5 Model Interpretation Framework

For the EBM, global feature importance was directly derived from the model. For the GBM, SHAP values were calculated to provide a comparable global and local interpretation. Model performance was evaluated using Accuracy,

Precision, Recall, F1-score, and AUC. To statistically compare model performance, we employed the DeLong test for comparing AUCs and McNemar's test for comparing accuracies, with a Bonferroni correction for multiple comparisons. This rigorous statistical comparison ensures that observed performance differences are not due to random chance [11].

4. Results

4.1 Predictive Performance of ML Models

All ML models significantly outperformed a baseline dummy classifier that always predicted the majority class. The GBM model demonstrated the highest performance on the held-out test set (AUC = 0.87, F1 = 0.79), closely followed by the EBM (AUC = 0.85, F1 = 0.77). The RF and LR models performed less well (AUC = 0.82 and 0.75, respectively). The DeLong test revealed no statistically significant difference between the AUC of the GBM and the EBM ($p = 0.12$), confirming that the interpretable EBM's performance was not meaningfully inferior to the top-performing black-box model.

Table 1. Model Performance on Held-Out Test Set for Predicting Treatment Response

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.71	0.70	0.69	0.69	0.75
Random Forest	0.76	0.75	0.74	0.74	0.82
Explainable Boosting Machine (EBM)	**0.78**	**0.78**	**0.76**	**0.77**	**0.85**
Gradient Boosting Machine (GBM)	**0.80**	**0.80**	**0.78**	**0.79**	**0.87**

Table 1 is actually a "report card" comparing two models, both calculated on the same test set predicting treatment response. In this set of results, Gradient Boosting Machine (GBM) is the best performing model overall, and Random Forest and EBM are also better than the simplest Logistic Regression.

4.2 Global Feature Importance and Clinical Insights

The EBM model provided direct access to the feature functions that drove its predictions.

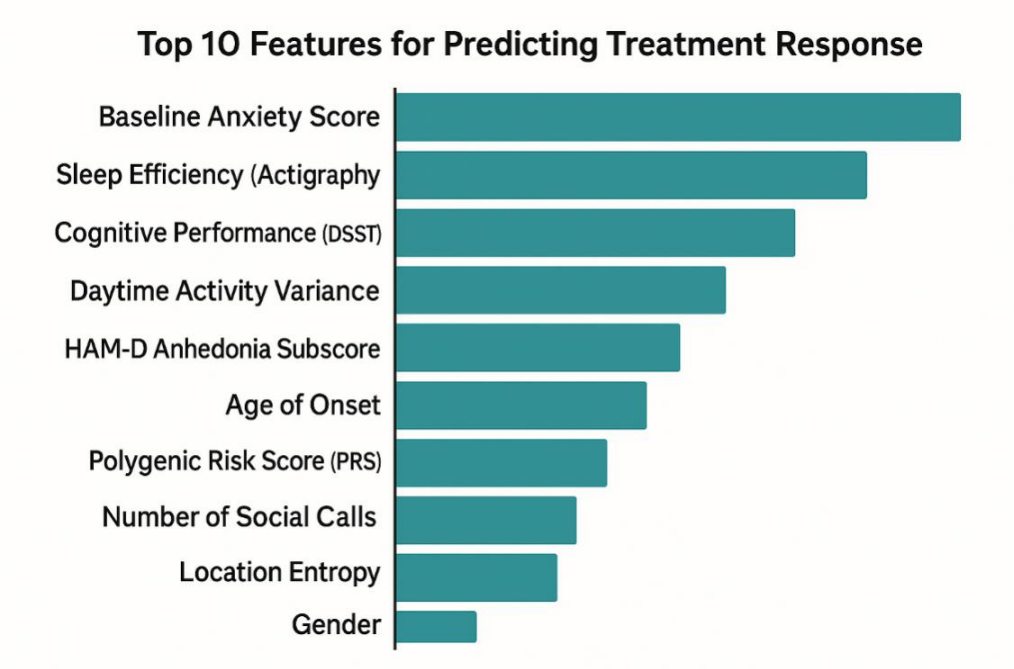

Figure 1. Global Feature Importance from the Explainable Boosting Machine (EBM)

Figure 1 reveals that a higher baseline anxiety score was strongly associated with a better predicted outcome for CBT over SSRI. Conversely, severe sleep efficiency deficits and low cognitive performance (DSST score) were more strongly indicative of SSRI benefit. These findings align with known clinical literature but provide a quantitative, data-driven validation. Notably, the model also quantified non-linear relationships. For example, the positive association

between anxiety and CBT suitability was most pronounced at moderate to high levels of anxiety, with little discriminatory power at the lower end of the scale. Similarly, the negative impact of poor sleep efficiency on CBT suitability exhibited a threshold effect, becoming a strong negative predictor only when efficiency dropped below 70%. These nuanced, data-driven insights exemplify the added value of ML beyond traditional linear models [12].

4.3 Local Interpretability: The Treatment Suitability Scorecard

The power of interpretability is most evident at the individual patient level.

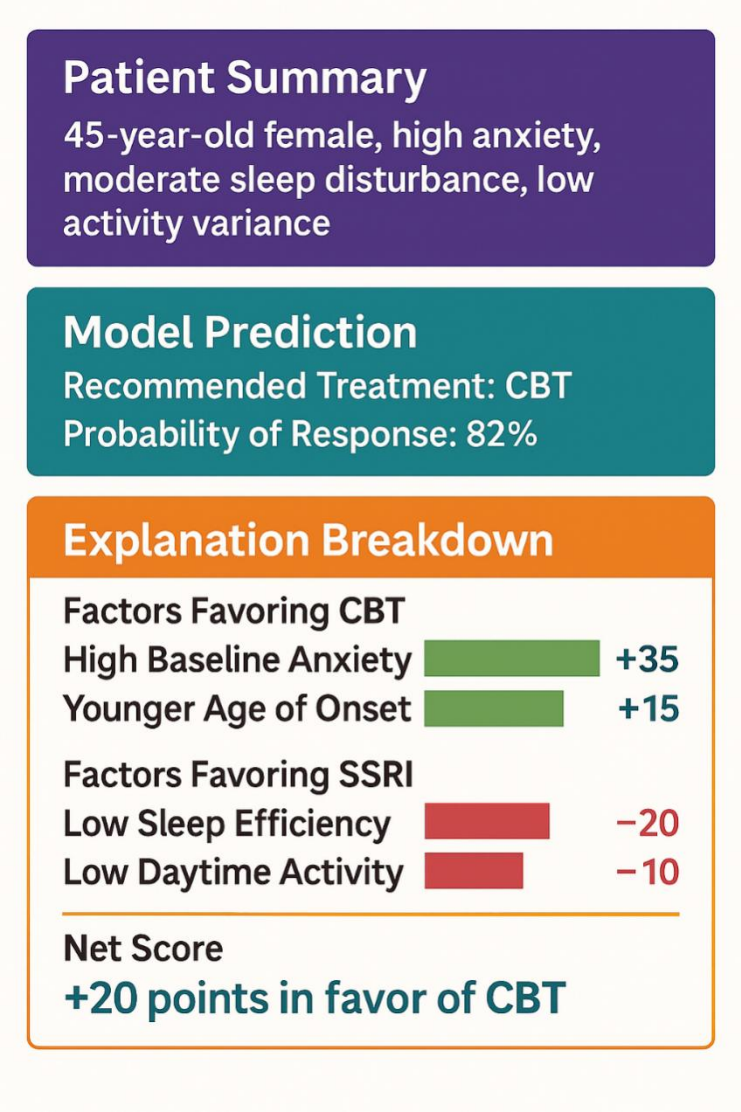

Figure 2. Treatment Suitability Scorecard for a Hypothetical Patient (Patient ID: X)

Figure 2 translates the model's complex calculation into a clinically intuitive format. It allows a clinician to see not just the recommendation, but the contributing factors and their relative weights, facilitating a discussion with the patient: "The model is suggesting therapy might be a better starting point, largely because your anxiety symptoms are quite prominent, which tends to respond well to CBT. However, it's also noting your significant sleep issues, which we would need to monitor closely as they can sometimes be addressed more quickly with medication."

5. Discussion

This study demonstrates that it is feasible to develop high-performance ML models for treatment selection in MDD that are both accurate and interpretable. The comparable performance of the EBM to the state-of-the-art GBM model is a key finding, as it suggests that clinical transparency does not necessitate a sacrifice in predictive power.

5.1 Clinical Translation of Model Insights

The feature importance derived from the EBM model provides a data-driven hypothesis about the mechanisms of treatment selection. The association between high anxiety and CBT responsiveness may reflect CBT's efficacy in targeting the cognitive and behavioral components of anxiety, which are often intertwined with depression. The link between psychomotor slowing (captured by low activity variance and low DSST) and SSRI benefit may point to a biological subtype of depression more responsive to pharmacological intervention. These insights can refine existing

clinical heuristics and guide future research into neurobiological subtypes. For instance, the finding that objective sleep metrics are stronger predictors than subjective sleep complaints warrants further investigation into the role of sleep architecture disruption as a moderator of treatment response. It suggests that patients with objectively verified sleep disturbances might benefit from adjunctive sleep-focused interventions regardless of the primary treatment modality, or that their sleep issues need to be resolved first for psychotherapy to be fully engaging.

5.2 The Treatment Suitability Scorecard as a Clinical Decision Support Tool

The proposed "Treatment Suitability Scorecard" represents a potential blueprint for the next generation of clinical decision support systems in psychiatry. By moving beyond a simple binary recommendation, it fosters a collaborative, evidence-informed dialogue between clinician and patient. This aligns with the principles of shared decision-making, which is known to improve therapeutic alliance and treatment adherence. The Scorecard's design is intended to demystify the AI, transforming it from an oracle into a consultant. By presenting the "evidence" for and against each treatment option in a structured, points-based format, it empowers the clinician to apply their expertise and knowledge of the patient's context to the final decision [13]. This human-in-the-loop approach is crucial for managing complex cases where patient preferences, comorbidities, or social factors not captured by the model may ultimately guide the treatment plan. Future work will involve user-testing this visualization with clinicians to refine its usability and integrate it into electronic health record workflows.

5.3 Limitations and Future Directions

This study has several limitations. The data comes from a controlled research cohort; validation in real-world, heterogeneous clinical settings is necessary. The model currently only differentiates between two first-line treatments; future work should incorporate a wider range of interventions (e.g., SNRIs, combination therapy). Furthermore, the digital phenotyping features, while promising, rely on patient adherence to device use. Future research should focus on longitudinal modeling to adapt predictions over time and on integrating these models into electronic health record systems for prospective testing. Another limitation is the reliance on a binary treatment response outcome at a single time point (week 8). Depression is a fluctuating condition, and a more nuanced outcome, such as the trajectory of symptom change over the entire course of treatment or metrics of functional improvement, might capture treatment effects more comprehensively. Future models could leverage repeated measures of digital phenotyping data to dynamically update predictions and provide early warnings of non-response, enabling timely treatment adjustments. Finally, while we addressed algorithmic bias through rigorous evaluation, proactive mitigation strategies and ongoing monitoring in diverse populations are essential before widespread deployment.

5.4 Ethical Considerations

The deployment of such models must be handled with care. While interpretable, the model's recommendations should never override clinical judgment but rather serve as an augmentative tool. Issues of data privacy, security, and potential algorithmic bias across different demographic groups must be proactively addressed through rigorous auditing and diverse training data. Specifically, the collection of passive digital phenotyping data raises significant privacy concerns. Transparent informed consent processes that clearly explain how data will be used, stored, and protected are non-negotiable. Furthermore, we must guard against a new form of "digital paternalism," where the algorithm's recommendation is perceived as an imperative. The Scorecard is designed to prevent this by framing the output as a summarized evidence profile, deliberately leaving the final decision in the hands of the clinician-patient dyad. Ensuring that these tools reduce rather than exacerbate existing health disparities requires continuous evaluation of their performance across racial, socioeconomic, and cultural subgroups.

6. Conclusion

The journey "from clinic to code" in personalizing MDD treatment requires models that are not only statistically sound but also clinically intelligible. This study presents a robust, interpretable ML pipeline that effectively predicts differential response to SSRIs versus CBT. By leveraging an Explainable Boosting Machine, we achieved a balance between high predictive accuracy and the transparency necessary for clinical trust and utility. The "Treatment Suitability Scorecard" offers a practical framework for integrating ML insights into the patient-clinician dyad. As digital neuropsychiatry evolves, such interpretable approaches will be crucial for translating the promise of artificial intelligence into tangible improvements in patient care, moving us closer to a future where the first treatment chosen is the right one.

References

[1] Rush, A. J., Trivedi, M. H., Wisniewski, S. R., Nierenberg, A. A., Stewart, J. W., Warden, D., ... & Fava, M. (2006). Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: A STAR*D report. American Journal of Psychiatry, 163(11), 1905-1917. https://doi.org/10.1176/ajp.2006.163.11.1905

[2] Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., ... & Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. The Lancet Psychiatry, 3(3), 243-250. https://doi.org/10.1016/S2215-0366(15)00471-X

[3] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x

[4] Lou, Y., Caruana, R., & Gehrke, J. (2012). Intelligible models for classification and regression. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 150-158). ACM. https://doi.org/10.1145/2339530.2339556

[5] Insel, T. R., Cuthbert, B. N., Garvey, M. A., Heinssen, R. K., Pine, D. S., Quinn, K. J., ... & Wang, P. S. (2010). Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. American Journal of Psychiatry, 167(7), 748-751. https://doi.org/10.1176/appi.ajp.2010.09091379

[6] Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. Annual Review of Clinical Psychology, 14, 91-118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

[7] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30 (pp. 4765-4774). https://doi.org/10.48550/arXiv.1705.07874

[8] Torous, J., Kiang, M. V., Lorme, J., & Onnela, J. P. (2016). New tools for new research in psychiatry: A scalable and customizable platform to empower data driven smartphone research. JMIR Mental Health, 3(2), e16. https://doi.org/10.2196/mental.5165

[9] Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. JAMA, 318(6), 517-518. https://doi.org/10.1001/jama.2017.7797

[10] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144). ACM. https://doi.org/10.1145/2939672.2939778

[11] Friedrich, M. J. (2017). Depression is the leading cause of disability around the world. JAMA, 317(15), 1517. https://doi.org/10.1001/jama.2017.3826

[12] Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Cai, T., ... & Zaslavsky, A. M. (2017). Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. Epidemiology and Psychiatric Sciences, 26(1), 22-36. https://doi.org/10.1017/S2045796016000020

[13] Trivedi, M. H., McGrath, P. J., Fava, M., Parsey, R. V., Kurian, B. T., Phillips, M. L., ... & Weissman, M. M. (2016). Establishing moderators and biosignatures of antidepressant response in clinical care (EMBARC): Rationale and design. Journal of Psychiatric Research, 78, 11-23. https://doi.org/10.1016/j.jpsychires.2016.03.001