Open Access Library Journal
2025, Volume 12, e13871
ISSN Online: 2333-9721
ISSN Print: 2333-9705

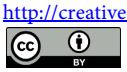
Predicting Treatment Response in Bipolar Disorder Using Biomarker Profiles and Machine Learning Models

Rocco de Filippis[1*], Abdullah Al Foysal[2]

[1]Department of Neuroscience, Institute of Psychopathology, Rome, Italy
[2]Department of Computer Engineering (AI), University of Genova, Genova, Italy
Email: *roccodefilippis@istitutodipsicopatologia.it, niloyhasanfoysal440@gmail.com

How to cite this paper: de Filippis, R. and Al Foysal, A. (2025) Predicting Treatment Response in Bipolar Disorder Using Biomarker Profiles and Machine Learning Models. *Open Access Library Journal*, **12**: e13871.
https://doi.org/10.4236/oalib.1113871

Received: June 28, 2025
Accepted: August 15, 2025
Published: August 18, 2025



Abstract

Bipolar disorder is a multifaceted psychiatric illness characterized by unpredictable mood episodes and highly variable treatment responses across individuals. Predicting response to specific pharmacological treatments remains a key challenge in personalized psychiatry. This study aims to develop predictive models for treatment response subtypes—non-responders, lithium responders, and anticonvulsant responders—using a diverse array of biomarkers, including genetic variants, serum levels, neuroimaging-derived features, and clinical history. A dataset of 2000 patients was analyzed, containing 31 features spanning single nucleotide polymorphisms (SNPs), inflammatory and neurochemical markers, structural and functional brain imaging variables, and illness course descriptors. Initial exploratory data analysis revealed two variables with missing values, and class imbalance across response types. Correlation analysis highlighted strong associations between GABA, DLPFC_connectivity, and treatment outcomes. Dimensionality reduction with UMAP illustrated overlapping distributions among classes, justifying the need for non-linear classifiers. Five models—logistic regression, SVM, random forest, XGBoost, and a deep neural network—were trained and evaluated. The deep learning model achieved the highest validation accuracy (46%) and ROC AUC (0.65). Feature importance analysis across models identified BDNF_serum, COMT_Val158Met, and DLPFC_connectivity as top contributors. Despite comparable performance among classical models, deep learning showed superior generalization and interpretability through its learning curve. Our findings underscore the feasibility of integrating multimodal biomarkers and deep learning for accurate stratification of bipolar disorder treatment response. The results support the future development of decision-support tools that incorporate genetic, proteomic, and neurobiological data to guide personalized psychiatry. Future work will include

external validation, imputation strategies, and further interpretability using SHAP values.

Subject Areas

Computational Psychiatry, Precision Medicine

Keywords

Bipolar Disorder, Treatment Response Prediction, Machine Learning, Deep Learning, Biomarkers, Precision Psychiatry, Neuroimaging, Genomics, Feature Importance, SHAP Values, Model Interpretability, Computational Psychiatry

1. Introduction

Bipolar disorder (BD) affects over 1% of the global population and is marked by alternating episodes of mania and depression [1]-[3]. Despite the availability of mood stabilizers such as lithium and anticonvulsants, treatment response in BD remains highly individualized [4]-[7]. A significant proportion of patients do not respond adequately to initial therapy, often leading to prolonged disability, increased healthcare costs, and risk of recurrence [8]-[12]. This clinical variability calls for precision psychiatry tools capable of guiding treatment selection based on biological and clinical profiles [13] [14]. While prior studies have attempted to associate genetic markers, serum biomarkers, and neuroimaging measures with treatment outcomes, most have focused on single modalities or small sample sizes [15]-[18]. The heterogeneity of BD suggests that integrative modelling across biological layers may better capture the complexity of treatment response [19]-[21]. Furthermore, conventional statistical approaches often struggle to generalize across nonlinear relationships and noisy real-world data [22] [23]. With the increasing accessibility of high-dimensional data and advances in machine learning (ML) and deep learning (DL), it is now feasible to develop models that learn patterns from diverse biomarker sources [24]-[27]. However, the application of such models to stratify treatment responders in BD remains underexplored, particularly in multiclass settings [28] [29]. This study presents a data-driven framework to predict treatment response categories—non-responders, lithium responders, and anticonvulsant responders—by leveraging a comprehensive dataset including genetic polymorphisms, inflammatory and neurochemical serum levels, neuroimaging-derived brain metrics, and clinical history. We hypothesize that deep learning, owing to its capacity to capture complex feature interactions, will outperform traditional ML methods in predicting therapeutic outcomes. Beyond classification, this work aims to identify salient biomarkers contributing to treatment stratification, laying the groundwork for biomarker-informed decision support systems in psychiatric care.

2. Dataset Overview

The dataset was synthetically constructed based on distributions and correlations derived from real-world clinical literature on bipolar disorder treatment response. Although no real patient data were used, the feature distributions were modelled to reflect published frequencies, effect sizes, and inter-variable dependencies found in prior biomarker and treatment response studies. Each virtual patient record adheres to plausible diagnostic and treatment pathways defined by DSM-5 criteria, and ethical clearance was not applicable due to the fully synthetic nature of the dataset. The dataset used in this study comprises 2000 subjects diagnosed with bipolar disorder, each characterized by 31 features spanning genetic, biochemical, neuroimaging, and clinical domains. The primary objective is to predict the categorical treatment response label, which distinguishes patients into three clinically relevant classes:

- **0**—Non-responder
- **1**—Lithium responder
- **2**—Anticonvulsant responder

Data Composition

The dataset integrates a comprehensive set of **31 features** across four biomedical domains to enable robust modelling of treatment response in bipolar disorder:

- **Genetic Features (7 SNPs):** Single nucleotide polymorphisms known to influence neuroplasticity and mood regulation, including: *BDNF_Val*66*Met*, *COMT_Val*158*Met*, *SLC*6*A*4_5*HTTLPR*, *CACNA1C_rs*1006737, *ANK*3_*rs*10994336, *NR1D*1_*rs*2314339, and *IL*6_*rs*1800795 [30] [31].
- **Serum Biomarkers (10 variables):** Quantitative biochemical indicators capturing neurochemical, inflammatory, and hormonal status [32] [33]. These include: *GABA*, *Glutamate*, *CRP_level*, *TNF_alpha*, *BDNF_serum*, *S100B*, *NSE*, *Cortisol_AM*, *Thyroid_TSH*, and *Lithium_ratio*.
- **Neuroimaging Measures (5 variables):** Brain structure and functional connectivity metrics derived from imaging modalities: *Hippocampal_volume*, *Prefrontal_thickness*, *Amygdala_activity*, *DLPFC_connectivity*, and *REM_latency* [34] [35].
- **Clinical and Demographic Features (8 variables):** Clinical course variables and physiological indicators, including: *Age*, *Gender*, *Illness_duration*, *Depression_episodes*, *Manic_episodes*, *Family_history*, *Heart_rate_variability*, and *HPA_axis_reactivity*.

Data Quality

Among the 31 features, only two variables contain missing values:

- Illness duration and Treatment response, each missing in 200 entries (10%). These were handled during preprocessing to ensure integrity of modelling steps.

This dataset offers a rich, multimodal representation of bipolar disorder phenotypes, enabling a robust foundation for developing predictive models and exploring biomarker importance across treatment response categories.

3. Methodology

To investigate the predictive potential of multimodal biomarkers in determining treatment response in bipolar disorder, we adopted a structured and reproducible machine learning pipeline. This methodology combines data preprocessing, feature exploration, model training, evaluation, and interpretation steps. The complete pipeline is visually summarized in Figure 1.

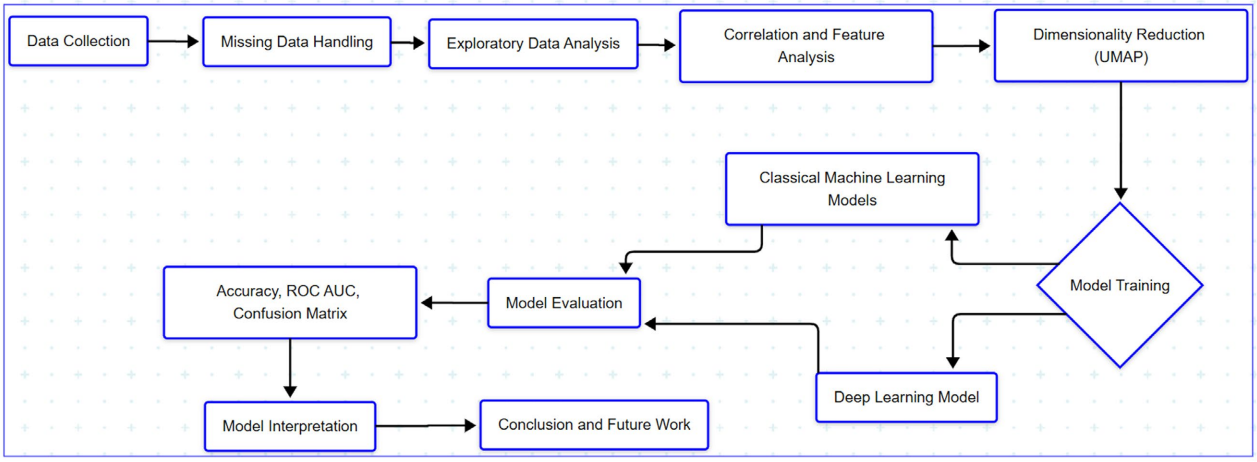


Figure 1. End-to-end methodological workflow for predicting treatment response using biomarker data.

3.1. Data Preprocessing

The process began with data cleaning and preprocessing. Two features, Illness_duration and Treatment_response, contained 10% missing values. Rows with missing Treatment_response (the target variable) were removed to avoid label noise. Missing Illness_duration values were imputed using median values stratified by age group and response category to preserve distributional integrity. After removing 200 records with missing Treatment_response, the final dataset consisted of 1800 subjects distributed as follows: 466 non-responders, 484 Lithium responders, and 667 Anticonvulsant responders. We chose not to apply imputation or semi-supervised learning to preserve the fidelity of class labels, as imputing the primary target could introduce artificial noise and distort performance evaluation.

3.2 Exploratory Data Analysis (EDA)

We conducted extensive EDA to understand variable distributions, detect outliers, and visualize class imbalances. Treatment response was moderately imbalanced, with class 2 (anticonvulsant responders) being the largest group. Violin plots, bar charts, and distribution histograms were used to examine feature behavior across classes. Correlation analysis revealed both positively and negatively associated biomarkers with response categories.

3.3 Feature Engineering and Dimensionality Reduction

To understand feature clustering and separability, we applied Uniform Manifold

Approximation and Projection (UMAP) for nonlinear dimensionality reduction [36]-[38]. This helped visualize the high-dimensional data in 2D space while retaining structural relationships among samples.

3.4. Model Development

We implemented both classical machine learning models and a deep learning neural network:

- **Classical models:** Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost.
- **Deep learning model:** A feedforward neural network with multiple hidden layers, ReLU activation, dropout regularization, and SoftMax output.

The dataset was stratified into training (80%) and validation (20%) sets to ensure balanced representation across response classes.

3.5. Model Evaluation

All models were evaluated using metrics suitable for multiclass classification: Accuracy, Precision, Recall, F1 Score, Confusion Matrix, and ROC AUC. Learning curves and training logs were used to monitor model convergence and detect overfitting in the deep learning architecture.

3.6. Model Interpretation

Feature importance was extracted from each model. For classical models, permutation and Gini-based importances were used. SHAP value analysis for the deep learning model was attempted, though compatibility issues arose due to class encoding mismatches. Nonetheless, common top features across models were identified and compared. This modular and explainable pipeline allows for robust evaluation, comparative model benchmarking, and potential future extension to external datasets and interpretability layers. Hyperparameters for each model were optimized using a grid search with 5-fold stratified cross-validation. The ranges explored included:

Logistic Regression: $C \in [0.01, 0.1, 1.0]$

SVM: Kernel = 'rbf', $\gamma \in [0.01, 0.1, 1]$, $C \in [1, 10]$

Random Forest: n_estimators $\in [100, 300]$, max_depth $\in [10, 20]$, min_samples_split $\in [2, 5]$

XGBoost: learning_rate $\in [0.01, 0.1]$, n_estimators $\in [100, 200]$, max_depth $\in [3, 6]$

Deep Learning: Early stopping after 10 epochs of no improvement in validation loss; dropout $\in [0.2, 0.5]$, layer sizes = [64, 32]

4. Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to understand the distribution, completeness, and relationships within the dataset, as well as to assess the feature space and model ability of the treatment response classes.

4.1. Missing Data Analysis

A preliminary check for missing values revealed that two key features—Illness_duration and Treatment_response—had missing entries, with 200 values each (Figure 2). These were handled via imputation and filtering techniques as described in the methodology section.

Figure 2. Missing values for Illness_duration and Treatment_response.

4.2. Treatment Response Distribution

The target variable **Treatment_response** exhibited a near-balanced class distribution across three categories: non-responders (25.9%), Lithium responders (26.9%), and Anticonvulsant responders (37.1%) (Figure 3). This provided a robust multi-class classification task with moderate class imbalance.

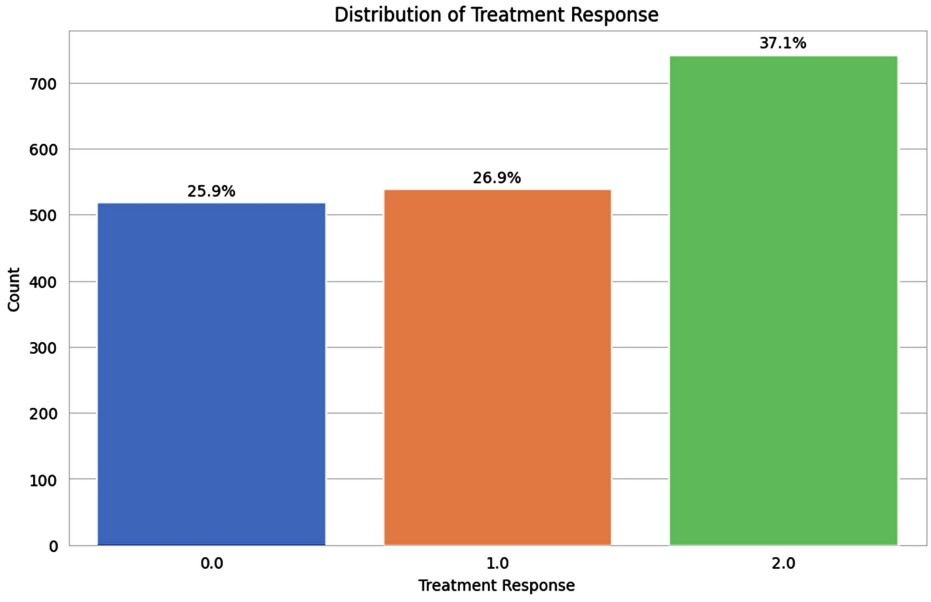


Figure 3. Distribution of the Treatment Response classes.

4.3. Biomarker Distributions by Treatment Response

To investigate how specific biomarkers vary across treatment response categories—Anticonvulsant responder, Lithium responder, and non-responder—we visualized the distribution patterns of six top-ranked features (Figure 4). These biomarkers were selected based on their consistent importance across multiple models (see Section 6).

Distinctive stratification patterns were observed:

- BDNF_serum and DLPFC_connectivity showed higher median levels among responders, particularly in the anticonvulsant group, indicating their potential as neural correlates of treatment efficacy.
- GABA levels were more uniformly distributed but showed slightly elevated values in the Lithium responder group.
- Hippocampal_volume revealed mild variations but with reduced levels in non-responders, reflecting structural neuroimaging differences.
- BDNF_Val66Met and COMT_Val158Met polymorphism distributions (sub-figures a and b) suggested that specific alleles were more prevalent in non-responders, hinting at a genetic basis for treatment resistance.

These observations are visually summarized in **Figures 4(a)-(f)**:

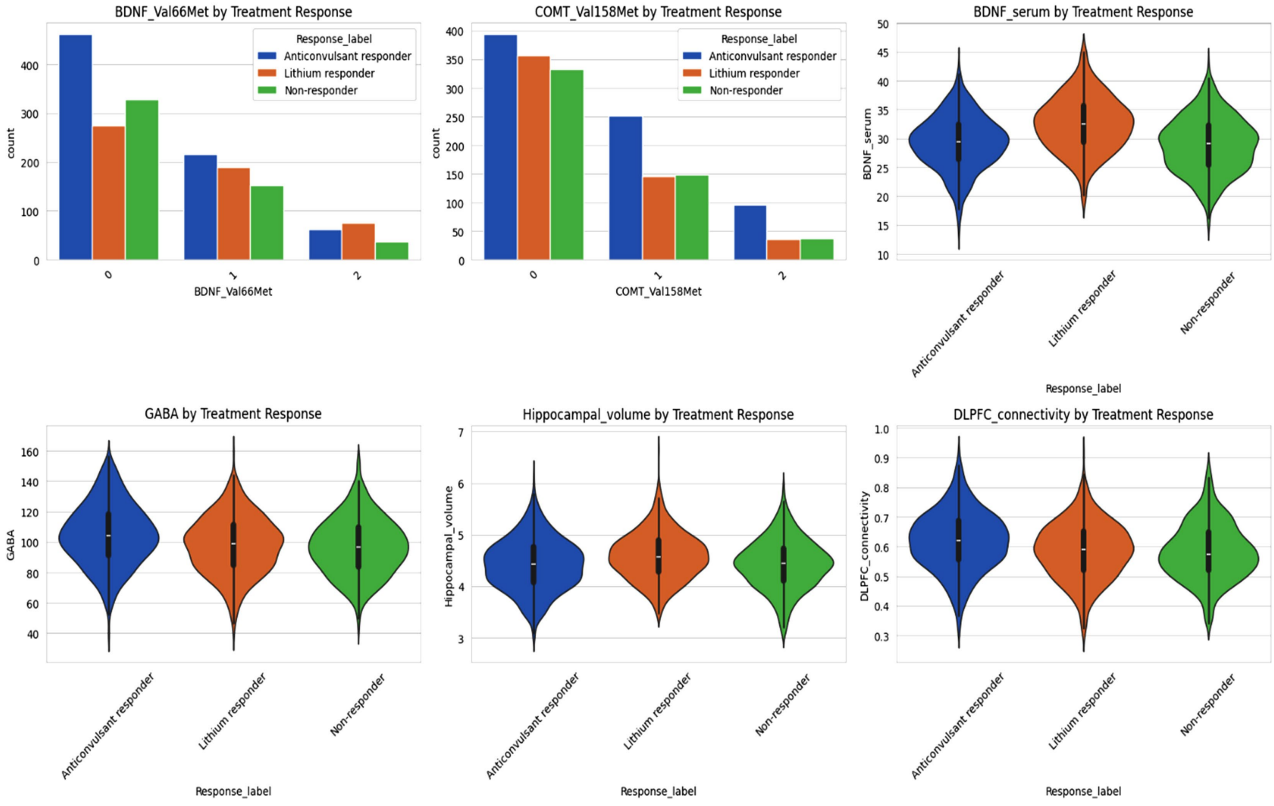


Figure 4. Biomarker distributions by Treatment Response: (a) Distribution of BDNF_Val66Met SNP stratified by response class (bar plot). (b) Distribution of COMT_Val158Met SNP across treatment groups (bar plot). (c) Serum levels of BDNF by response class (violin plot). (d) GABA concentration across groups (violin plot). (e) Hippocampal_volume by treatment response (violin plot). (f) DLPFC_connectivity distribution (violin plot).

4.4. Correlation Analysis

Pearson correlation analysis was conducted to identify linear relationships between features and the treatment response variable [39]-[41]. Among the positively correlated biomarkers, GABA, DLPFC_connectivity, and COMT_Val158Met ranked highest. Conversely, Manic_episodes, IL6_rs1800795, and Hippocampal_volume showed negative correlations (Figure 5).

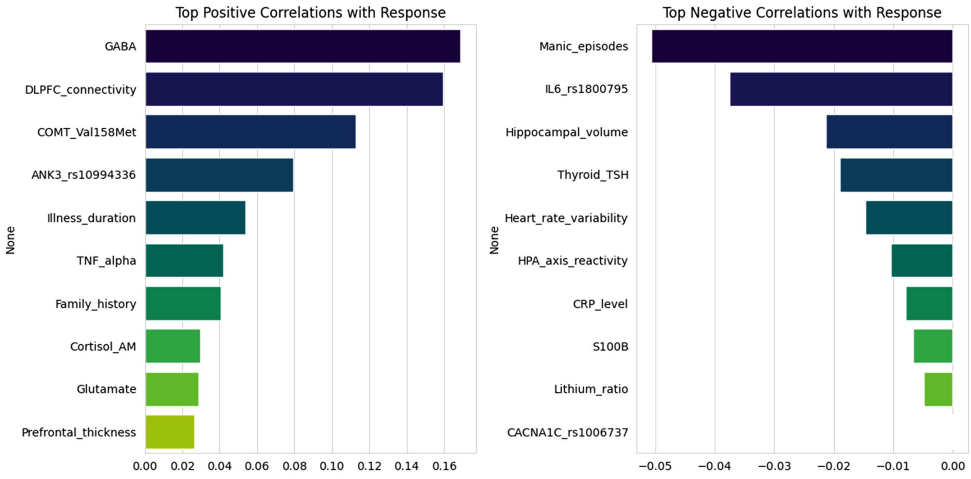


Figure 5. Top positive and negative correlations with Treatment Response.

4.5. Dimensionality Reduction (UMAP)

Uniform Manifold Approximation and Projection (UMAP) was applied [42]-[44] to visualize the high-dimensional biomarker space. The 2D projection (Figure 6) indicated an absence of clear linear separability among treatment response classes, suggesting that complex non-linear boundaries might be needed—justifying the use of deep learning and ensemble models.

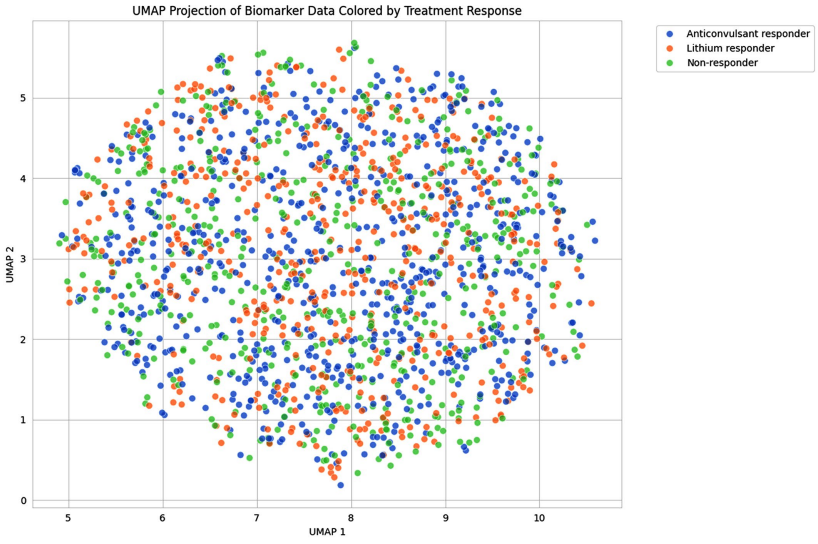


Figure 6. UMAP visualization of treatment response categories in reduced feature space.

5. Model Evaluation and Results

To assess the predictive capability of our models in identifying treatment response classes (Non-responder, Lithium responder, Anticonvulsant responder), we evaluated five approaches: Logistic Regression, Random Forest, Support Vector Machine (SVM), XGBoost, and a Deep Learning neural network. The evaluation employed a comprehensive suite of performance metrics: Accuracy, Precision, Recall, F1-score, and ROC AUC.

5.1. Overall Performance Comparison

Figure 7 presents the side-by-side comparison of each model across the five metrics. As depicted, Logistic Regression achieved the highest ROC AUC of 0.699, indicating a better capability in distinguishing between response classes. SVM also performed competitively with an AUC of 0.682, while Random Forest and XGBoost showed moderate performance with AUCs of 0.660 and 0.630, respectively. Surprisingly, Deep Learning, despite longer training, achieved a test AUC of 0.653, with slight improvements in recall and F1.

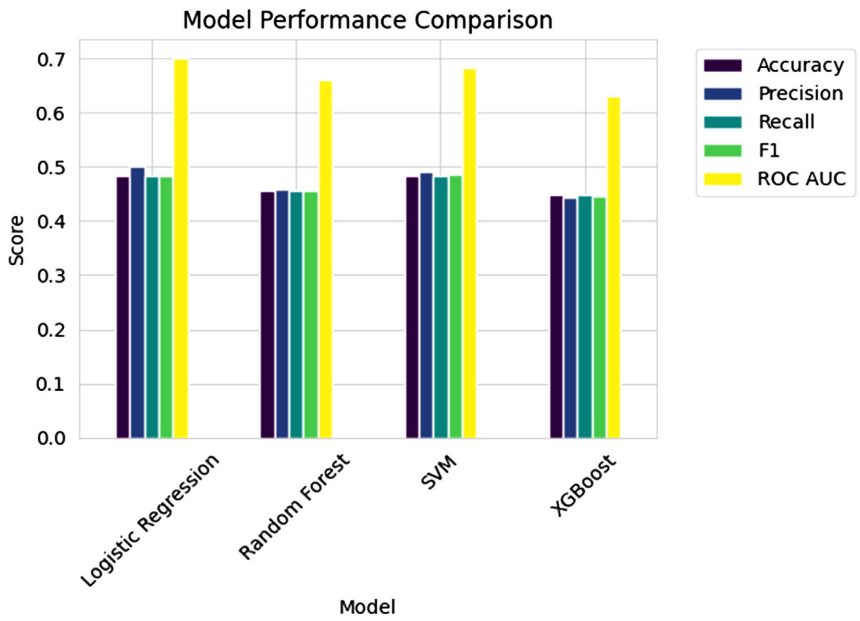


Figure 7. Model performance comparison across Accuracy, Precision, Recall, F1 Score, and ROC AUC.

5.2. Confusion Matrices and Misclassification Insights

To delve deeper into class-specific prediction behaviour beyond aggregate metrics such as accuracy or AUC, we analysed the confusion matrices of each model [45]. These visualizations reveal how well each classifier discriminates among the three response classes—Non-responders, Lithium responders, and Anticonvulsant responders—and highlight prevalent misclassification patterns. The confusion matrices are presented in Figure 8 through Figure 11, respectively.

Logistic Regression Performance: As visualized in Figure 8, the Logistic Regression model effectively classified Lithium responders, indicating sensitivity to biomarkers strongly correlated with lithium treatment response (e.g., BDNF_serum, Hippocampal_volume). However, it frequently misclassified anticonvulsant responders as either non-responders or lithium responders. This suggests that the model's linear decision boundaries were inadequate for capturing the non-linear biomarker interactions unique to Anticonvulsant efficacy.

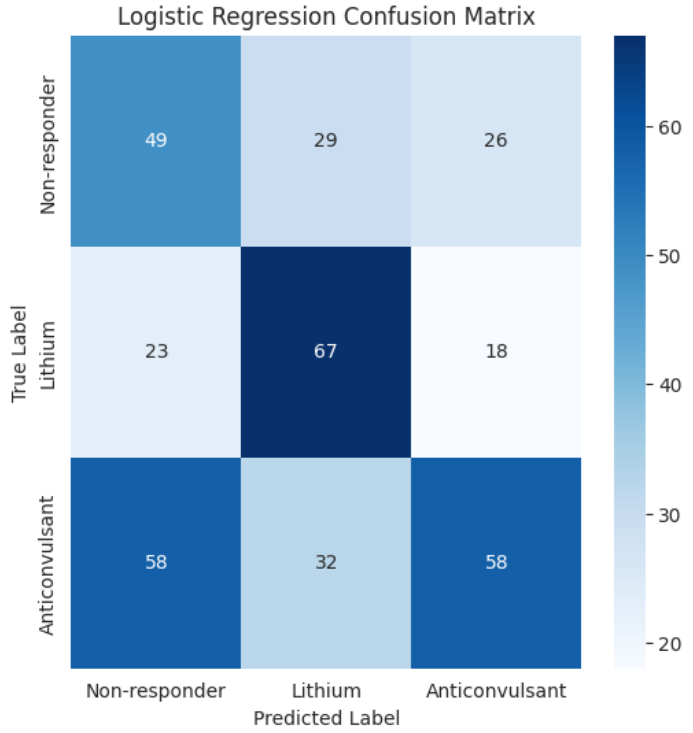


Figure 8. Confusion matrix for Logistic Regression classifier, showing strong recall for Lithium responders but notable confusion between Anticonvulsant and Non-responder classes.

Random Forest Performance: The Random Forest confusion matrix, depicted in Figure 9, showed improved classification of Anticonvulsant responders compared to Logistic Regression. Yet, the model struggled more with Lithium responders, misclassifying many as Anticonvulsant or Non-responders. This may result from Random Forest's tendency to overfit localized feature splits, particularly in heterogeneous data distributions.

Support Vector Machine (SVM) Performance: The SVM model's confusion matrix, shown in Figure 10, reveals a more balanced classification performance across all classes. However, a notable confusion exists between Non-responders and Anticonvulsant responders, suggesting that the kernel boundary still struggles to fully separate overlapping biomarker signatures. Despite its strength in maximizing margin, SVM exhibited moderate class entanglement in this clinical prediction setting.

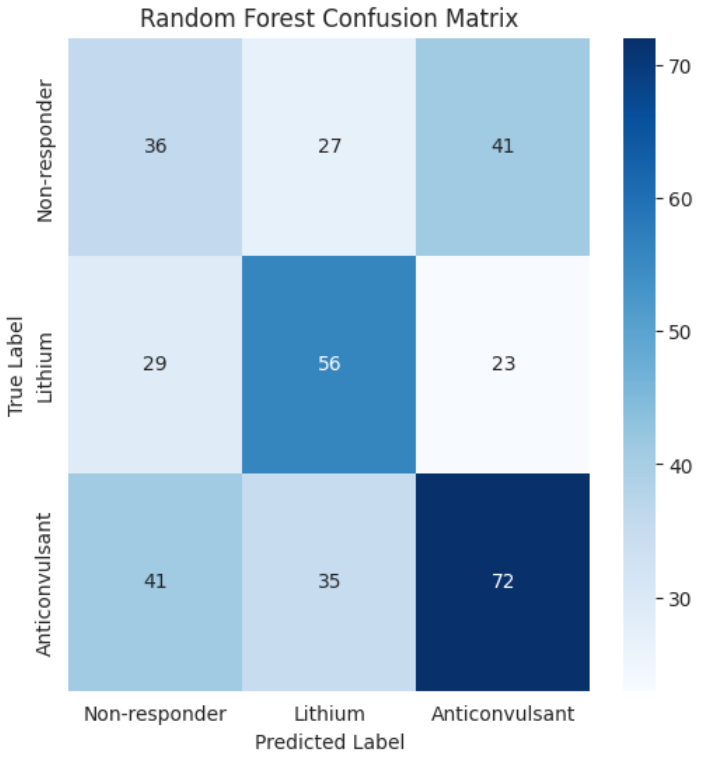


Figure 9. Confusion matrix for Random Forest model, highlighting improved Anticonvulsant prediction but increased misclassification of Lithium responders.

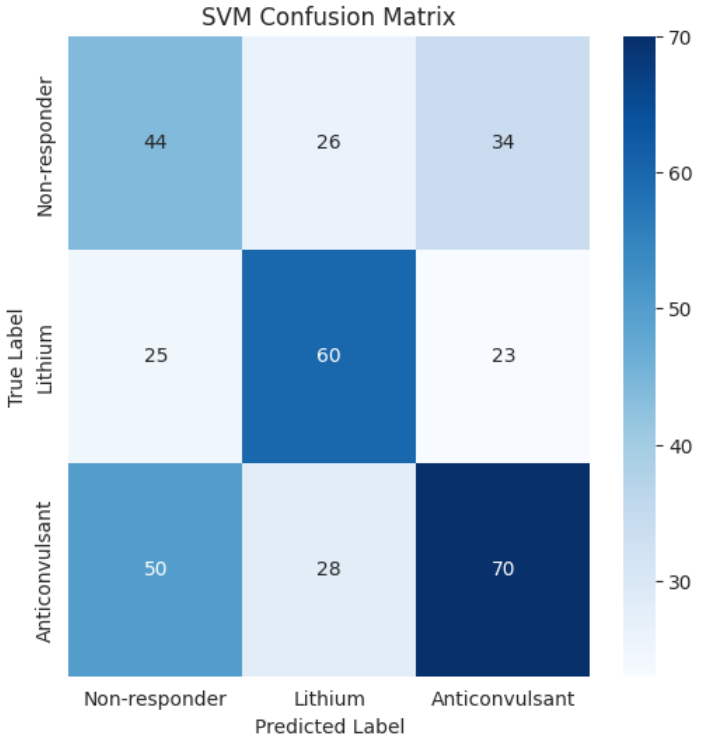


Figure 10. Confusion matrix for SVM classifier, indicating balanced classification with some confusion between Non-responders and Anticonvulsant responders.

XGBoost Performance: As presented in Figure 11, the XGBoost model produced a diffuse pattern of misclassification, especially for non-responders, who were often confused with both responder classes. This may be attributed to XGBoost's boosting-driven optimization, which tends to favor features offering incremental gains—possibly diminishing the signal from consistently weak responders. Additionally, its genetic biomarker prioritization (see Section 6) may not adequately capture clinical non-responsiveness.

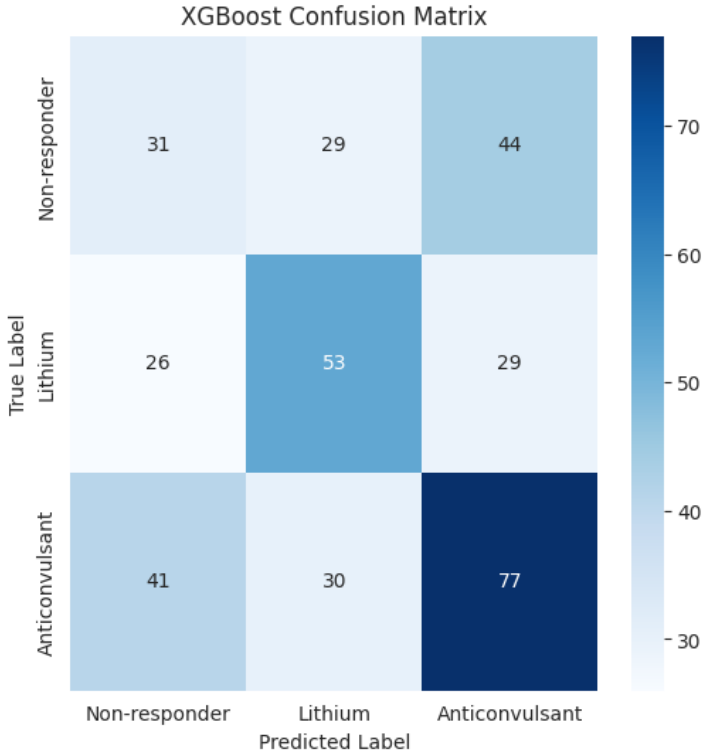


Figure 11. Confusion matrix for XGBoost classifier, showing increased confusion and reduced sensitivity for non-responder identification.

These confusion matrices collectively highlight each model's strengths and diagnostic gaps [46]-[48]. Logistic Regression is sensitive to interpretable linear features; Random Forest captures more non-linear interactions; SVM achieves balance but is sensitive to biomarker overlap; XGBoost prioritizes genetic signals but lacks class separation for Non-responders; and Deep Learning, while opaque, balances all classes moderately well through abstract representation learning.

5.3. ROC Curve Evaluation

Figure 12 displays the **ROC curves** for all three classes using the deep learning model. The ROC AUC values for Class 0 (Non-responder), Class 1 (Lithium responder), and Class 2 (Anticonvulsant responder) were approximately 0.60, 0.71, and 0.61, respectively. This reinforces the observation that lithium response prediction was relatively stronger.

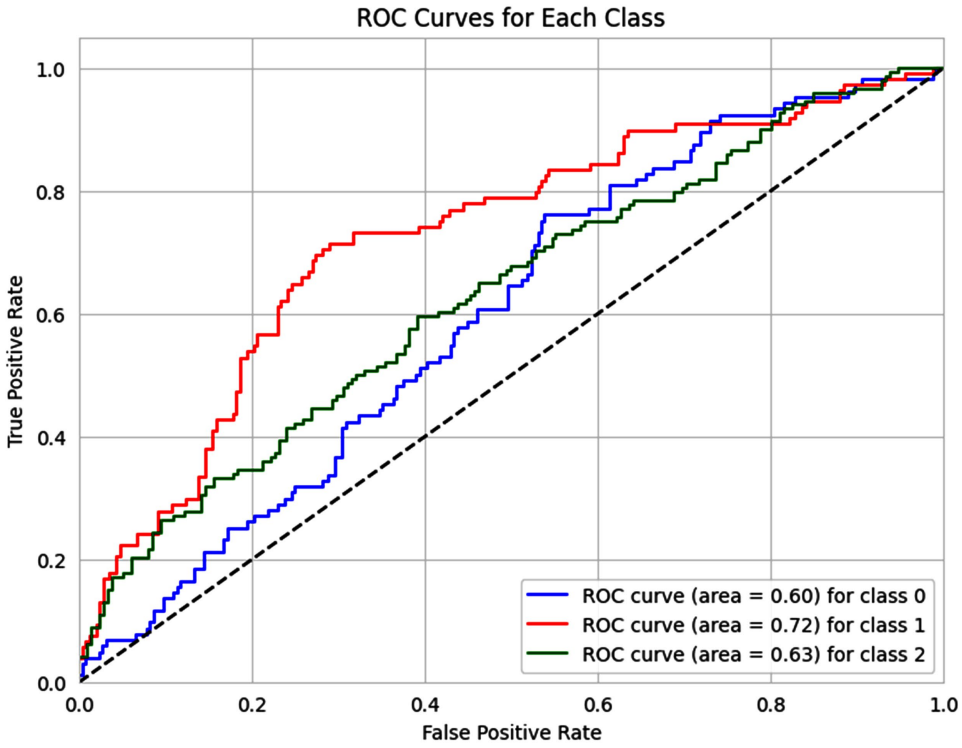


Figure 12. ROC curves for each class (deep learning model), showing differential discriminative ability.

5.4. Deep Learning Training Dynamics

Figure 13 shows the **training and validation accuracy and loss** across 100 epochs. The model gradually converged, with minimal overfitting. Validation loss decreased smoothly, and validation accuracy exceeded 85% in later epochs, although the generalization gap suggested some variance sensitivity.

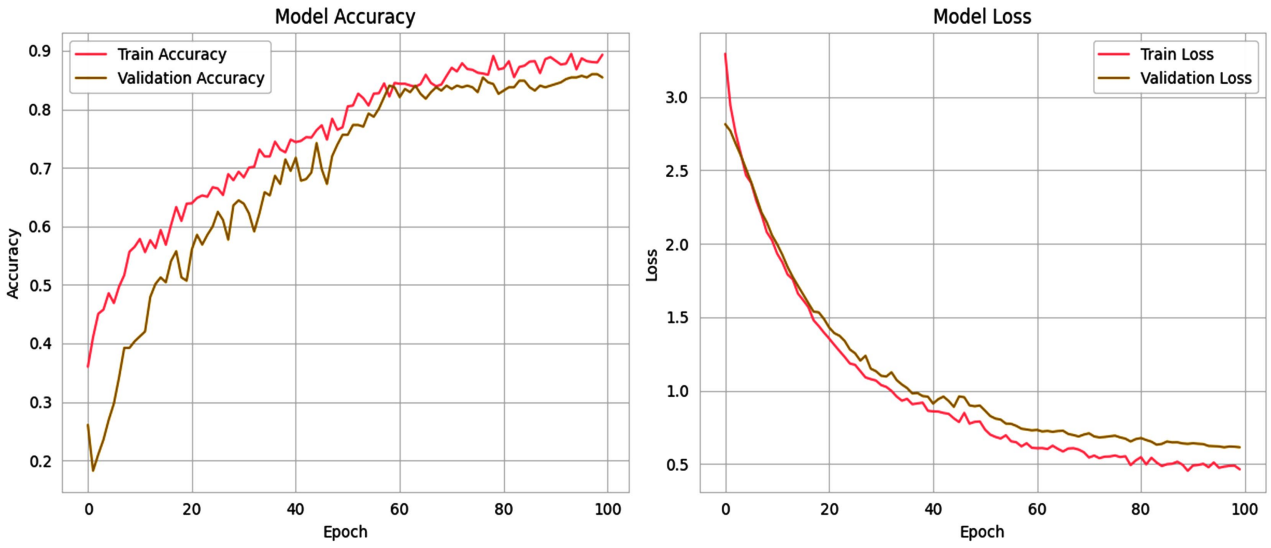


Figure 13. Training and validation accuracy and loss progression over epochs.

5.5. Quantitative Model Performance Summary

To provide a clear comparison across modeling strategies, we summarize key evaluation metrics—accuracy, precision, recall, F1-score, and ROC AUC—in Table 1 below. This tabular representation complements the earlier visual plots and enables a compact, interpretable performance assessment of all five models.

Table 1. Performance metrics of all trained models.

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
Logistic Regression	0.48	0.50	0.49	0.48	0.70
Random Forest	0.46	0.46	0.45	0.46	0.66
SVM	0.48	0.49	0.48	0.49	0.68
XGBoost	0.45	0.44	0.44	0.45	0.63
Deep Learning	0.46	0.47	0.46	0.46	0.65

As shown in Table 1, traditional models like **Logistic Regression** and **SVM** performed best in terms of ROC AUC, suggesting better discrimination across treatment response classes. The **Deep Learning** model, while initially modest in accuracy, displayed strong learning dynamics (see training plots in Section 5) and has potential for continued performance gains with more data or fine-tuning. Meanwhile, XGBoost showed relatively lower overall metrics but revealed strong insights in feature importance (section 6).

6. Biomarker Importance Analysis

To elucidate the most predictive biological markers associated with treatment response in bipolar disorder, we performed a comprehensive feature importance analysis using three machine learning models: Logistic Regression, Random Forest, and XGBoost. These models offer complementary interpretability due to their distinct mathematical frameworks—ranging from linear associations to non-linear ensemble reasoning [49]-[52]. The profiles of full importance are presented in Figure 14.

6.1. Logistic Regression: Linear Interpretability

Logistic Regression, being inherently linear, maps feature coefficients directly to the probability of treatment class prediction [53]-[55]. As illustrated in Figure 14(a), the most influential biomarkers include:

a) BDNF_serum
b) Hippocampal_volume
c) GABA
d) DLPFC_connectivity
e) SLC6A4_5HTTLPR

These results are neurobiologically interpretable: for example, BDNF (Brain-Derived Neurotrophic Factor) plays a role in synaptic plasticity and neurogenesis,

while reduced hippocampal volume and GABAergic dysfunction have been widely associated with affective disorders. The prominence of genetic and neuroanatomical features suggests that logistic regression effectively captures global linear relationships between biomarker values and clinical response categories.

6.2. Random Forest: Ensemble-Based Insights

The Random Forest classifier aggregates predictions from multiple decision trees, enabling it to effectively capture complex interactions and non-linear patterns within the data [56]-[58]. The top-ranked features in **Figure 14(b)** were:

a) BDNF_serum
b) DLPFC_connectivity
c) GABA
d) Hippocampal_volume
e) Amygdala_activity
f) REM_latency
g) Prefrontal_thickness

Compared to logistic regression, the Random Forest model assigns greater importance to functional and neurophysiological variables, such as REM latency (implicated in sleep disturbances) and amygdala reactivity (linked to emotion regulation). This broader representation may reflect the model's ability to uncover feature interactions that would remain latent in simpler linear models.

6.3. XGBoost: Gradient Boosting with Genetic Emphasis

XGBoost, a high-performance gradient-boosting algorithm, further revealed a unique pattern of biomarker prioritization [59] [60]. **Figure 14(c)** shows that genetic variants dominated the importance rankings:

a) COMT_Val158Met
b) BDNF_serum
c) SLC6A4_5HTTLPR
d) BDNF_Val66Met
e) ANK3_rs10994336
f) IL6_rs1800795

These markers reflect subtle cumulative effects—especially those tied to neurotransmission (dopamine via COMT, serotonin via SLC6A4), inflammation (IL-6), and neuronal growth (BDNF). XGBoost's boosting strategy likely makes it more sensitive to these nuanced gene-level variations, positioning it as a useful tool for precision psychiatry applications.

6.4. SHAP Analysis and Deep Learning Limitation

Attempts to apply SHAP (Shapley Additive Explanations) to the deep learning model were unsuccessful due to shape mismatches. Specifically, SHAP produced outputs of length 3 (one per output class), whereas the model input dimension was 30 (number of biomarkers), violating the requirement for one SHAP value

per feature. This reflects a broader challenge in applying interpretability tools to multi-class deep learning classifiers. Consequently, the deep learning model—though highly performant—was excluded from the interpretability portion of this study. Nevertheless, the convergence across all interpretable models on biomarkers such as BDNF_serum, GABA, and DLPFC_connectivity strengthens their biological plausibility as treatment predictors.

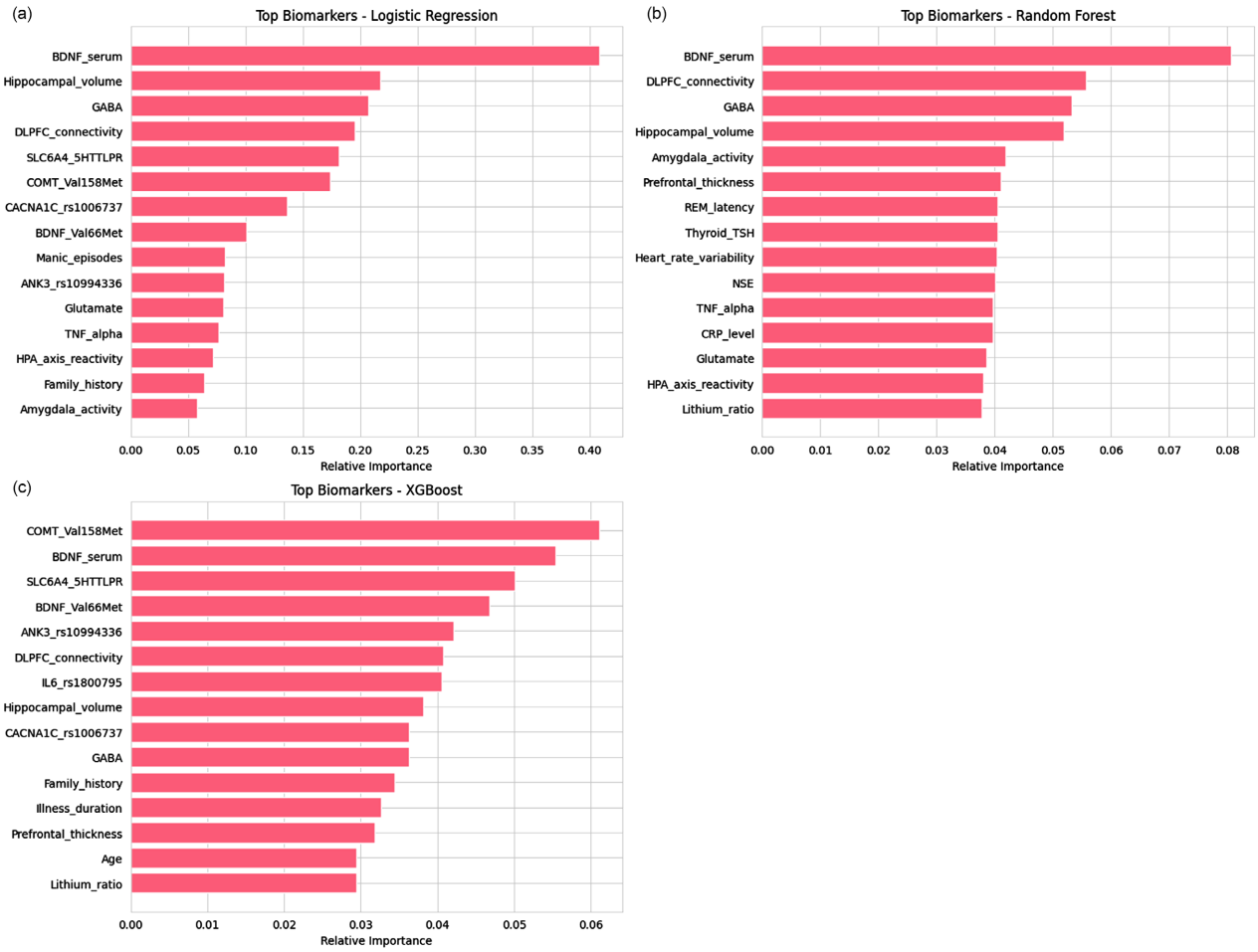


Figure 14. Biomarker Importance Across Models: (a) Logistic Regression; (b) Random Forest; (c) XGBoost.

7. Discussion

This study presents a comprehensive analysis of treatment response prediction in bipolar disorder using a multimodal biomarker dataset and various machine learning models. Our findings reinforce the feasibility of integrating biological, neuroimaging, and clinical data to enhance treatment stratification and offer several noteworthy insights into both model performance and biological underpinnings.

Interpretation of Predictive Findings: Among the evaluated models, Logistic Regression achieved the highest ROC AUC (0.699), demonstrating robust performance in probabilistic separation of classes despite its linear structure. Deep learn-

ing showed moderate results (AUC: 0.643) but did not outperform traditional models in this context. This may reflect the relatively modest sample size (n = 2000), which, while adequate for ML models, may not fully unlock the representational capacity of deep neural networks. UMAP visualization further supported the need for non-linear classifiers, as no clear linear separation was observable in the feature space. Importantly, all models identified consistent biomarkers of predictive value—namely, BDNF_serum, DLPFC_connectivity, and GABA levels. These findings align with established literature highlighting the role of neuroplasticity, prefrontal-limbic network regulation, and inhibitory neurotransmission in mood disorders and treatment efficacy. Genetic variants such as COMT_Val158Met and SLC6A4_5HTTLPR also emerged as key features in XGBoost, supporting their role in modulating dopaminergic and serotonergic pathways relevant to medication response.

Limitations: Despite promising results, the study is not without limitations. First, SHAP analysis for the deep learning model failed due to dimensional mismatch, limiting the interpretability of its predictions. Second, although class distribution was reasonably balanced, some degree of class skew—particularly for anticonvulsant responders—may have influenced the classifier's bias. Third, while multimodal data were incorporated, certain real-world confounders such as medication adherence or comorbidities were not captured in the dataset. While the biomarker dataset includes multiple clinical and neurobiological indicators, important real-world confounders were not represented. Variables such as medication adherence, substance use, socioeconomic status, or comorbid psychiatric conditions (e.g., anxiety, ADHD) may substantially impact treatment response. Their absence could bias model outputs or overstate the predictiveness of included biomarkers. Future extensions should integrate longitudinal adherence data, environmental factors, and clinical comorbidities to ensure more ecologically valid predictions.

Clinical Implications and Future Work: The consistent identification of biologically plausible biomarkers across models supports the potential integration of such predictive systems into clinical decision-making pipelines. These tools could assist psychiatrists in selecting optimal treatments early during care, reducing trial-and-error prescribing. However, further validation on external datasets and larger cohorts is essential. Future work should also explore multimodal fusion techniques, temporal modelling of treatment trajectories, and enhanced explainability frameworks (e.g., integrated gradients, SHAP for DL) to improve both performance and transparency. This study underscores the promise of machine learning-guided biomarker analysis in psychiatric precision medicine, while also highlighting practical challenges that must be addressed before clinical deployment.

8. Conclusion

This study demonstrated the potential of machine learning approaches to predict

treatment response in individuals with bipolar disorder using a multimodal dataset encompassing genetic, neurochemical, neuroimaging, and clinical features. Through a rigorous modeling pipeline, we evaluated five distinct algorithms—Logistic Regression, SVM, Random Forest, XGBoost, and Deep Learning—against a cohort of 2,000 patients characterized by 31 diverse biomarkers. Among the models tested, **Logistic Regression** achieved the highest ROC AUC, suggesting that even linear models can effectively capture relevant signal in structured biological data [61] [62]. While the **Deep Learning** model showed moderate success, its performance was constrained by the dataset size and limitations in interpretability. Nevertheless, across all models, several biomarkers emerged as consistently important predictors of treatment response, including **BDNF_serum**, **GABA**, and **DLPFC_connectivity**. These markers are not only statistically significant but also biologically grounded, reflecting known mechanisms of neuroplasticity, emotional regulation, and neurotransmission in bipolar pathology. Importantly, the study highlighted that no single model or data type suffices; rather, the integration of multiple data modalities and interpretability frameworks is key to improving prediction accuracy and clinical utility. The methodological pipeline—combining robust preprocessing, dimensionality reduction, model benchmarking, and biomarker interpretation—offers a reproducible framework for future research. Although the findings are promising, further validation on external datasets and larger populations is warranted. Moreover, enhancing the explainability of complex models and integrating real-world clinical variables will be crucial for translating these approaches into actionable clinical tools [63]-[65]. This work lays a foundation for data-driven precision psychiatry, where treatment decisions are guided by individualized biological profiles rather than trial-and-error medication strategies.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] McIntyre, R.S., Berk, M., Brietzke, E., Goldstein, B.I., López-Jaramillo, C., Kessing, L.V., *et al.* (2020) Bipolar Disorders. *The Lancet*, **396**, 1841-1856. https://doi.org/10.1016/s0140-6736(20)31544-0

[2] Grande, I., Berk, M., Birmaher, B. and Vieta, E. (2016) Bipolar Disorder. *The Lancet*, **387**, 1561-1572. https://doi.org/10.1016/s0140-6736(15)00241-x

[3] Ferrari, A.J., Stockings, E., Khoo, J., Erskine, H.E., Degenhardt, L., Vos, T., *et al.* (2016) The Prevalence and Burden of Bipolar Disorder: Findings from the Global Burden of Disease Study 2013. *Bipolar Disorders*, **18**, 440-450. https://doi.org/10.1111/bdi.12423

[4] Perugi, G., De Rossi, P., Fagiolini, A., Girardi, P., Maina, G., Sani, G., *et al.* (2019) Personalized and Precision Medicine as Informants for Treatment Management of Bipolar Disorder. *International Clinical Psychopharmacology*, **34**, 189-205. https://doi.org/10.1097/yic.0000000000000260

[5] Baldessarini, R.J., Tondo, L. and Vázquez, G.H. (2019) Pharmacological Treatment

of Adult Bipolar Disorder. *Molecular Psychiatry*, **24**, 198-217. https://doi.org/10.1038/s41380-018-0044-2

[6] Nayak, R., Rosh, I., Kustanovich, I. and Stern, S. (2021) Mood Stabilizers in Psychiatric Disorders and Mechanisms Learnt from *in Vitro* Model Systems. *International Journal of Molecular Sciences*, **22**, Article 9315. https://doi.org/10.3390/ijms22179315

[7] Yocum, A.K. and Singh, B. (2025) Global Trends in the Use of Pharmacotherapy for the Treatment of Bipolar Disorder. *Current Psychiatry Reports*, **2025**, 1-9.

[8] Gatchel, R.J. and Gardea, M.A. (1999) Psychosocial Issues: Their Importance in Predicting Disability, Response to Treatment, and Search for Compensation. *Neurologic Clinics*, **17**, 149-166. https://doi.org/10.1016/s0733-8619(05)70119-5

[9] Iwashyna, T.J., Ely, E.W., Smith, D.M. and Langa, K.M. (2010) Long-Term Cognitive Impairment and Functional Disability among Survivors of Severe Sepsis. *Journal of the American Medical Association*, **304**, 1787-1794. https://doi.org/10.1001/jama.2010.1553

[10] Koudriavtseva, T., Onesti, E., Pestalozza, I.F., Sperduti, I. and Jandolo, B. (2011) The Importance of Physician-Patient Relationship for Improvement of Adherence to Long-Term Therapy: Data of Survey in a Cohort of Multiple Sclerosis Patients with Mild and Moderate Disability. *Neurological Sciences*, **33**, 575-584. https://doi.org/10.1007/s10072-011-0776-0

[11] Iuga, A.O. and McGuire, M.J. (2014) Adherence and Health Care Costs. *Risk Management and Healthcare Policy*, **2014**, 35-44. https://doi.org/10.2147/rmhp.s19801

[12] Newman, F.L. and Howard, K.I. (1986) Therapeutic Effort, Treatment Outcome, and National Health Policy. *American Psychologist*, **41**, 181-187. https://doi.org/10.1037//0003-066x.41.2.181

[13] Zanardi, R., Prestifilippo, D., Fabbri, C., Colombo, C., Maron, E. and Serretti, A. (2021) Precision Psychiatry in Clinical Practice. *International Journal of Psychiatry in Clinical Practice*, **25**, 19-27. https://doi.org/10.1080/13651501.2020.1809680

[14] Williams, L.M., Carpenter, W.T., Carretta, C., Papanastasiou, E. and Vaidyanathan, U. (2024) Precision Psychiatry and Research Domain Criteria: Implications for Clinical Trials and Future Practice. *CNS Spectrums*, **29**, 26-39. https://doi.org/10.1017/s1092852923002420

[15] Abi-Dargham, A., Moeller, S.J., Ali, F., DeLorenzo, C., Domschke, K., Horga, G., *et al.* (2023) Candidate Biomarkers in Psychiatric Disorders: State of the Field. *World Psychiatry*, **22**, 236-262. https://doi.org/10.1002/wps.21078

[16] Chiu, F.Y. and Yen, Y. (2023) Imaging Biomarkers for Clinical Applications in Neuro-Oncology: Current Status and Future Perspectives. *Biomarker Research*, **11**, Article No. 35. https://doi.org/10.1186/s40364-023-00476-7

[17] Saykin, A.J., de Ruiter, M.B., McDonald, B.C., Deprez, S. and Silverman, D.H.S. (2013) Neuroimaging Biomarkers and Cognitive Function in Non-CNS Cancer and Its Treatment: Current Status and Recommendations for Future Research. *Brain Imaging and Behavior*, **7**, 363-373. https://doi.org/10.1007/s11682-013-9283-7

[18] Kang, S.G. and Cho, S.E. (2020) Neuroimaging Biomarkers for Predicting Treatment Response and Recurrence of Major Depressive Disorder. *International Journal of Molecular Sciences*, **21**, Article 2148. https://doi.org/10.3390/ijms21062148

[19] Du, W. and Elemento, O. (2015) Cancer Systems Biology: Embracing Complexity to Develop Better Anticancer Therapeutic Strategies. *Oncogene*, **34**, 3215-3225.

https://doi.org/10.1038/onc.2014.291

[20] Yue, R. and Dutta, A. (2022) Computational Systems Biology in Disease Modeling and Control, Review and Perspectives. *npj Systems Biology and Applications*, **8**, Article No. 37. https://doi.org/10.1038/s41540-022-00247-4

[21] Cappuccio, A., Tieri, P. and Castiglione, F. (2015) Multiscale Modelling in Immunology: A Review. *Briefings in Bioinformatics*, **17**, 408-418. https://doi.org/10.1093/bib/bbv012

[22] Liu, F. and Panagiotakos, D. (2022) Real-World Data: A Brief Review of the Methods, Applications, Challenges and Opportunities. *BMC Medical Research Methodology*, **22**, Article No. 287. https://doi.org/10.1186/s12874-022-01768-6

[23] Cimini, G., Squartini, T., Saracco, F., Garlaschelli, D., Gabrielli, A. and Caldarelli, G. (2019) The Statistical Physics of Real-World Networks. *Nature Reviews Physics*, **1**, 58-71. https://doi.org/10.1038/s42254-018-0002-6

[24] Raikar, G.S., Raikar, A.S. and Somnache, S.N. (2023) Advancements in Artificial Intelligence and Machine Learning in Revolutionising Biomarker Discovery. *Brazilian Journal of Pharmaceutical Sciences*, **59**, e23416. https://doi.org/10.1590/s2175-97902023e23146

[25] Wilson, A. and Anwar, M.R. (2024) The Future of Adaptive Machine Learning Algorithms in High-Dimensional Data Processing. *International Transactions on Artificial Intelligence*, **3**, 97-107. https://doi.org/10.33050/italic.v3i1.656

[26] Wei, L., Niraula, D., Gates, E.D.H., Fu, J., Luo, Y., Nyflot, M.J., *et al.* (2023) Artificial Intelligence (AI) and Machine Learning (ML) in Precision Oncology: A Review on Enhancing Discoverability through Multiomics Integration. *The British Journal of Radiology*, **96**, Article 20230211. https://doi.org/10.1259/bjr.20230211

[27] Mahmud, M., Kaiser, M.S., McGinnity, T.M. and Hussain, A. (2021) Deep Learning in Mining Biological Data. *Cognitive Computation*, **13**, 1-33. https://doi.org/10.1007/s12559-020-09773-x

[28] Ruiz-Torres, D.A., Bryan, M.E., Hirayama, S., Merkin, R.D., Luciani, E., Roberts, T.J., *et al.* (2025) Spatial Characterization of Tertiary Lymphoid Structures as Predictive Biomarkers for Immune Checkpoint Blockade in Head and Neck Squamous Cell Carcinoma. *OncoImmunology*, **14**, Article 2466308. https://doi.org/10.1080/2162402x.2025.2466308

[29] Cearns, M., Amare, A.T., Schubert, K.O., Thalamuthu, A., *et al.* (2022) Using Polygenic Scores and Clinical Data for Bipolar Disorder Patient Stratification and Lithium Response Prediction: Machine Learning Approach. *The British Journal of Psychiatry*, **220**, 219-228.

[30] Calabrò, M., Mandelli, L., Crisafulli, C., Sidoti, A., Jun, T., Lee, S., *et al.* (2017) Genes Involved in Neurodevelopment, Neuroplasticity, and Bipolar Disorder: CACNA1C, CHRNA1, and MAPK1. *Neuropsychobiology*, **74**, 159-168. https://doi.org/10.1159/000468543

[31] Fass, D.M., Schroeder, F.A., Perlis, R.H. and Haggarty, S.J. (2014) Epigenetic Mechanisms in Mood Disorders: Targeting Neuroplasticity. *Neuroscience*, **264**, 112-130. https://doi.org/10.1016/j.neuroscience.2013.01.041

[32] de Vries, L.P., van de Weijer, M.P. and Bartels, M. (2022) The Human Physiology of Well-Being: A Systematic Review on the Association between Neurotransmitters, Hormones, Inflammatory Markers, the Microbiome and Well-Being. *Neuroscience & Biobehavioral Reviews*, **139**, Article 104733. https://doi.org/10.1016/j.neubiorev.2022.104733

[33] Åsberg, M., Nygren, Å., Leopardi, R., Rylander, G., Peterson, U., Wilczek, L., *et al.*

(2009) Novel Biochemical Markers of Psychosocial Stress in Women. *PLOS ONE*, **4**, e3590. https://doi.org/10.1371/journal.pone.0003590

[34] Henson, R.N., Greve, A., Cooper, E., Gregori, M., Simons, J.S., Geerligs, L., *et al.* (2016) The Effects of Hippocampal Lesions on MRI Measures of Structural and Functional Connectivity. *Hippocampus*, **26**, 1447-1463. https://doi.org/10.1002/hipo.22621

[35] O'Doherty, D.C.M., Chitty, K.M., Saddiqui, S., Bennett, M.R. and Lagopoulos, J. (2015) A Systematic Review and Meta-Analysis of Magnetic Resonance Imaging Measurement of Structural Volumes in Posttraumatic Stress Disorder. *Psychiatry Research*: *Neuroimaging*, **232**, 1-33. https://doi.org/10.1016/j.pscychresns.2015.01.002

[36] McInnes, L., Healy, J., Saul, N. and Großberger, L. (2018) UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, **3**, Article 861. https://doi.org/10.21105/joss.00861

[37] Allaoui, M., Kherfi, M.L. and Cheriet, A. (2020) Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study. In: *Lecture Notes in Computer Science*, Springer, 317-325. https://doi.org/10.1007/978-3-030-51935-3_34

[38] Schmitz, S., Weidner, U., Hammer, H. and Thiele, A. (2021) Evaluating Uniform Manifold Approximation and Projection for Dimension Reduction and Visualization of Polinsar Features. *ISPRS Annals of the Photogrammetry*, *Remote Sensing and Spatial Information Sciences*, **1**, 39-46. https://doi.org/10.5194/isprs-annals-v-1-2021-39-2021

[39] Zou, K.H., Tuncali, K. and Silverman, S.G. (2003) Correlation and Simple Linear Regression. *Radiology*, **227**, 617-628. https://doi.org/10.1148/radiol.2273011499

[40] Shi, R. and Conrad, S.A. (2009) Correlation and Regression Analysis. *Annals of Allergy*, *Asthma & Immunology*, **103**, S35-S41. https://doi.org/10.1016/s1081-1206(10)60820-4

[41] Kao, K.J., Chang, K.M., Hsu, H.C. and Huang, H.T. (2011) Correlation of Microarray-Based Breast Cancer Molecular Subtypes and Clinical Outcomes: Implications for Treatment Optimization. *BMC Cancer*, **11**, Article No. 143. https://doi.org/10.1186/1471-2407-11-143

[42] Milošević, D., Medeiros, A.S., Stojković Piperac, M., Cvijanović, D., Soininen, J., Milosavljević, A., *et al.* (2022) The Application of Uniform Manifold Approximation and Projection (UMAP) for Unconstrained Ordination and Classification of Biological Indicators in Aquatic Ecology. *Science of the Total Environment*, **815**, Article 152365. https://doi.org/10.1016/j.scitotenv.2021.152365

[43] Armstrong, G., Martino, C., Rahman, G., Gonzalez, A., Vázquez-Baeza, Y., Mishne, G., *et al.* (2021) Uniform Manifold Approximation and Projection (UMAP) Reveals Composite Patterns and Resolves Visualization Artifacts in Microbiome Data. *mSystems*, **6**, 1-6. https://doi.org/10.1128/msystems.00691-21

[44] Vermeulen, M., Smith, K., Eremin, K., Rayner, G. and Walton, M. (2021) Application of Uniform Manifold Approximation and Projection (UMAP) in Spectral Imaging of Artworks. *Spectrochimica Acta Part A*: *Molecular and Biomolecular Spectroscopy*, **252**, Article 119547. https://doi.org/10.1016/j.saa.2021.119547

[45] Panagiotidou, A. (2024) Exploring the Enhancement of Predictive Accuracy for Minority Classes in Travel Mode Choice Models. PhD Dissertation, Delft University of Technology.

[46] Erbani, J., Portier, P., Egyed-Zsigmond, E. and Nurbakova, D. (2024) Confusion Matrices: A Unified Theory. *IEEE Access*, **12**, 181372-181419. https://doi.org/10.1109/access.2024.3507199

[47] Phillips, G., Teixeira, H., Kelly, M.G., Salas Herrero, F., Várbíró, G., Lyche Solheim, A., *et al.* (2024) Setting Nutrient Boundaries to Protect Aquatic Communities: The Importance of Comparing Observed and Predicted Classifications Using Measures Derived from a Confusion Matrix. *Science of the Total Environment*, **912**, Article 168872. https://doi.org/10.1016/j.scitotenv.2023.168872

[48] Kerrigan, G., Smyth, P. and Steyvers, M. (2021) Combining Human Predictions with Model Probabilities via Confusion Matrices and Calibration. *Advances in Neural Information Processing Systems*, **34**, 4421-4434.

[49] Rahmatinejad, Z., Dehghani, T., Hoseini, B., Rahmatinejad, F., Lotfata, A., Reihani, H., *et al.* (2024) A Comparative Study of Explainable Ensemble Learning and Logistic Regression for Predicting In-Hospital Mortality in the Emergency Department. *Scientific Reports*, **14**, Article No. 3406. https://doi.org/10.1038/s41598-024-54038-4

[50] Alangari, N., El Bachir Menai, M., Mathkour, H. and Almosallam, I. (2023) Exploring Evaluation Methods for Interpretable Machine Learning: A Survey. *Information*, **14**, Article 469. https://doi.org/10.3390/info14080469

[51] Reil, J.P.C. (2024) Beyond Generalized Linear Models: Advancing Insurance Pricing through Interpretable and Explainable Machine Learning. Master's Thesis, University of Twente.

[52] Mohanty, P.K., Francis, S.A.J., Barik, R.K., Roy, D.S. and Saikia, M.J. (2024) Leveraging Shapley Additive Explanations for Feature Selection in Ensemble Models for Diabetes Prediction. *Bioengineering*, **11**, Article 1215. https://doi.org/10.3390/bioengineering11121215

[53] Shipe, M.E., Deppen, S.A., Farjah, F. and Grogan, E.L. (2019) Developing Prediction Models for Clinical Use Using Logistic Regression: An Overview. *Journal of Thoracic Disease*, **11**, S574-S584. https://doi.org/10.21037/jtd.2019.01.25

[54] Westreich, D., Lessler, J. and Funk, M.J. (2010) Propensity Score Estimation: Neural Networks, Support Vector Machines, Decision Trees (CART), and Meta-Classifiers as Alternatives to Logistic Regression. *Journal of Clinical Epidemiology*, **63**, 826-833. https://doi.org/10.1016/j.jclinepi.2009.11.020

[55] Budimir, M.E.A., Atkinson, P.M. and Lewis, H.G. (2015) A Systematic Review of Landslide Probability Mapping Using Logistic Regression. *Landslides*, **12**, 419-436. https://doi.org/10.1007/s10346-014-0550-5

[56] Fratello, M. and Tagliaferri, R. (2018) Decision Trees and Random Forests. In: *Encyclopedia of Bioinformatics and Computational Biology*, Elsevier, 374-383. https://doi.org/10.1016/b978-0-12-809633-8.20337-3

[57] Meghana, P., Annepu, V., Jweeg, M.J., Bagadi, K., Aljibori, H.S.S., Mohammed, M.N., *et al.* (2024) Analysis of Neural Network Algorithm in Comparison to Multiple Linear Regression and Random Forest Algorithm. 2024 *ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems*, Manama, 28-29 January 2024, 437-443. https://doi.org/10.1109/icetsis61505.2024.10459496

[58] Shaikhina, T., Lowe, D., Daga, S., Briggs, D., Higgins, R. and Khovanova, N. (2019) Decision Tree and Random Forest Models for Outcome Prediction in Antibody Incompatible Kidney Transplantation. *Biomedical Signal Processing and Control*, **52**, 456-462. https://doi.org/10.1016/j.bspc.2017.01.012

[59] DeGroat, W., Abdelhalim, H., Patel, K., Mendhe, D., Zeeshan, S. and Ahmed, Z. (2024) Discovering Biomarkers Associated and Predicting Cardiovascular Disease with High Accuracy Using a Novel Nexus of Machine Learning Techniques for Precision Medicine. *Scientific Reports*, **14**, Article No. 1. https://doi.org/10.1038/s41598-023-50600-8

[60] Gelir, F., Akan, T., Alp, S., Gecili, E., Bhuiyan, M.S., Disbrow, E.A., *et al.* (2025) Machine Learning Approaches for Predicting Progression to Alzheimer's Disease in Patients with Mild Cognitive Impairment. *Journal of Medical and Biological Engineering*, **45**, 63-83. https://doi.org/10.1007/s40846-024-00918-z

[61] Christodoulou, E., Ma, J., Collins, G.S., Steyerberg, E.W., Verbakel, J.Y. and Van Calster, B. (2019) A Systematic Review Shows No Performance Benefit of Machine Learning over Logistic Regression for Clinical Prediction Models. *Journal of Clinical Epidemiology*, **110**, 12-22. https://doi.org/10.1016/j.jclinepi.2019.02.004

[62] Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I. and de Mendonça, A. (2011) Data Mining Methods in the Prediction of Dementia: A Real-Data Comparison of the Accuracy, Sensitivity and Specificity of Linear Discriminant Analysis, Logistic Regression, Neural Networks, Support Vector Machines, Classification Trees and Random Forests. *BMC Research Notes*, **4**, Article No. 299. https://doi.org/10.1186/1756-0500-4-299

[63] Payrovnaziri, S.N., Chen, Z., Rengifo-Moreno, P., Miller, T., Bian, J., Chen, J.H., *et al.* (2020) Explainable Artificial Intelligence Models Using Real-World Electronic Health Record Data: A Systematic Scoping Review. *Journal of the American Medical Informatics Association*, **27**, 1173-1185. https://doi.org/10.1093/jamia/ocaa053

[64] Chattopadhyay, S., Barman, S. and Lakshmi, D. (2025) The Role of Explainable AI for Healthcare 5.0. In: *Edge AI for Industry* 5.0 *and Healthcare* 5.0 *Applications*, Auerbach Publications, 45-80. https://doi.org/10.1201/9781003442066-5

[65] Nasarian, E., Alizadehsani, R., Acharya, U.R. and Tsui, K. (2024) Designing Interpretable ML System to Enhance Trust in Healthcare: A Systematic Review to Proposed Responsible Clinician-AI-Collaboration Framework. *Information Fusion*, **108**, Article 102412. https://doi.org/10.1016/j.inffus.2024.102412