



**Master's Study Program between Hellenic Open
University and Ionian University**

Bioinformatics and Neuroinformatics

Master Thesis

*"Application of Machine Learning Methods for the Diagnosis of
Mental Disorders"*

Author's Name & Surname

Anastasia Maria Vitsiou

Supervisor: Prof. Themis Exarchos

Patras, Greece, December, 2024

Intellectual Property

Theses / Dissertations remain the intellectual property of students (“authors/creators”), but in the context of open access policy they grant to the HOU a non-exclusive license to use the right of reproduction, customization, public lending, presentation to an audience and digital dissemination thereof internationally, in electronic form and by any means for teaching and research purposes, for no fee and throughout the duration of intellectual property rights. Free access to the full text for studying and reading does not in any way mean that the author/creator shall allocate his/her intellectual property rights, nor shall he/she allow the reproduction, republication, copy, storage, sale, commercial use, transmission, distribution, publication, execution, downloading, uploading, translating, modifying in any way, of any part or summary of the dissertation, without the explicit prior written consent of the author/creator. Creators retain all their moral and property rights.



Application of Machine Learning Methods for the Diagnosis of Mental Disorders

Author: Anastasia Maria Vitsiou

Supervising Committee

Supervisor: Themis Exarchos, Associate Professor, Dept of Informatics, Ionian
University

Co-Supervisor: Marios Krokidis, Assistant Professor, Dept of Informatics,
Ionian University

Patras, Greece,
December, 2024

Acknowledgments / Dedication

I would like to express my sincere gratitude to my supervisors for their invaluable guidance and encouragement throughout this journey. To my family, my unwavering safe harbour, thank you for your endless support and for always inspiring me to chase my dreams. To my friends, whose presence made every challenge lighter, I am deeply grateful. And to my husband, for his love, care, and constant motivation, I couldn't have done this without you.

Abstract

Schizophrenia (SCZ) and Bipolar Disorder (BD) are severe psychiatric disorders that significantly impact individuals' well-being and functionality. The accurate diagnosis of these conditions remains a challenge due to overlapping symptomatology, a lack of definitive biomarkers, and limitations in traditional classification systems. This study leverages machine learning (ML) techniques to enhance the diagnostic accuracy of SCZ and BD using gene expression data. An XGBoost classifier was employed to perform multiclass classification on an imbalanced dataset, distinguishing between SCZ, BD, and healthy controls. The study utilized a stratified 5-fold cross-validation approach to ensure robust evaluation, with performance metrics including accuracy, precision, recall, F1-score, and ROC-AUC.

The results demonstrated the model's effectiveness, achieving an accuracy of 92.87% on the imbalanced dataset and 90% on the balanced dataset. Feature importance analysis using SHAP identified 90 key genes associated with SCZ and BD, with enrichment analyses revealing significant biological pathways, including glutamate receptor signalling and neuropeptide regulation. The findings highlight the potential of ML-driven genetic analysis to provide more objective diagnostic tools, complementing traditional clinical assessments.

Keywords: Schizophrenia (SCZ), Bipolar Disorder (BD), Gene Expression Data, Machine Learning (ML), SHAP Analysis, Biomarkers for Psychiatric Disorders, Neurobiological Pathways

Εφαρμογή Μεθόδων Μηχανικής Μάθησης για τη Διάγνωση Ψυχικών Διαταραχών

Αναστασία Μαρία Βίτσιου

Περίληψη

Η σχιζοφρένεια και η διπολική διαταραχή είναι σοβαρές ψυχιατρικές παθήσεις που επηρεάζουν σημαντικά την ευημερία και τη λειτουργικότητα των ατόμων. Η ακριβής διάγνυσή τους παραμένει πρόκληση λόγω της επικάλυψης της συμπτωματολογίας, της έλλειψης καθοριστικών βιοδεικτών και των περιορισμών των παραδοσιακών συστημάτων ταξινόμησης. Η παρούσα μελέτη αξιοποιεί τεχνικές μηχανικής μάθησης για τη βελτίωση της διαγνωστικής ακρίβειας της σχιζοφρένειας και της διπολικής διαταραχής, χρησιμοποιώντας δεδομένα γονιδιακής έκφρασης. Ένας ταξινομητής XGBoost εφαρμόστηκε για την πολυκατηγορική ταξινόμηση σε ένα μη ισορροπημένο σύνολο δεδομένων, για τη διάκριση μεταξύ σχιζοφρένειας, διπολικής διαταραχής και υγιών ατόμων. Η μελέτη χρησιμοποίησε stratified 5-fold cross-validation για τη διασφάλιση αξιόπιστης αξιολόγησης, με δείκτες απόδοσης όπως η ακρίβεια (accuracy), η επανακλησιμότητα (recall), η ακρίβεια πρόβλεψης (precision), το F1-score και η ROC-AUC.

Τα αποτελέσματα ανέδειξαν την αποτελεσματικότητα του μοντέλου, επιτυγχάνοντας ακρίβεια 92,87% στο μη ισορροπημένο σύνολο δεδομένων και 90% στο ισορροπημένο. Η ανάλυση της συνεισφοράς των χαρακτηριστικών μέσω SHAP αποκάλυψε 90 βασικά γονίδια που σχετίζονται με τη σχιζοφρένεια και τη διπολική διαταραχή, ενώ οι αναλύσεις εμπλουτισμού ανέδειξαν σημαντικές βιολογικές οδούς, συμπεριλαμβανομένων της σηματοδότησης των υποδοχέων γλουταμινικού και της ρύθμισης των νευροπεπτιδίων. Τα ευρήματα αναδεικνύουν τη δυνατότητα συνδυασμού της γενετικής ανάλυσης με τεχνικές μηχανικής μάθησης για την ανάπτυξη αντικειμενικότερων διαγνωστικών εργαλείων, ενισχύοντας και συμπληρώνοντας τις παραδοσιακές κλινικές αξιολογήσεις.

Λέξεις – Κλειδιά: Σχιζοφρένεια, Διπολική Διαταραχή, Δεδομένα γονιδιακής έκφρασης, Μηχανική Μάθηση, Ανάλυση SHAP, Βιοδείκτες Ψυχιατρικών Διαταραχών, Νευροβιολογικές Οδοί.

Contents

1	Introduction	1
1.1	Statement of Purpose	1
1.2	Objectives	4
2	Theoretical Background	6
2.1	Machine Learning	6
2.1.1	Key Types of Machine Learning	6
2.1.2	Gradient Boosting Decision Trees	8
2.1.3	XGBoost: A Boosting Algorithm	10
2.2	Bipolar Disorder and Schizophrenia	11
2.2.1	Anatomical, Biological, Biochemical, and Genetic Perspectives on BP and SZ	11
3	Literature Review	13
3.1	Traditional Diagnostic Methods	13
3.2	Diagnostic Methods Using Machine Learning	13
4	Methodology	18
4.1	Dataset Description	18
4.1.1	Data Collection and Phenotype Definition	18
4.1.2	Dataset Details	19
4.1.3	Data Pre-processing Methods	19
4.2	Machine Learning Model	20
4.2.1	Training Methodology	20
4.2.2	Tools, Frameworks, and Setup	21
4.3	Evaluation Metrics	21
4.3.1	23
5	Results	25
5.1	Overview of Model's Performance	25
5.2	Insights gained from the results	31
6	Conclusion	38
6.1	Summary of Findings	38
6.2	Contributions	38
6.3	Implications	38
6.4	Limitations	39
6.5	Future Work	39

List of Figures

1.1	Bipolar Disorder prevalence, 2021[24].	3
1.2	Schizophrenia prevalence, 2021[25].	4
2.1	Key types of Machine Learning and algorithms[5]	6
2.2	Boosting Pseudo code[5]	8
2.3	The tree structure generated from the dataset used in this study, following training with a Decision Tree Classifier.	9
2.4	Operational framework of the Gradient Boosted Decision Trees (GBDT)	10
2.5	General architecture of XGBoost algorithm [1].	10
5.1	Confusion Matrix for the first approach	26
5.2	Receiver Operating Characteristic (ROC) curve for the first approach	27
5.3	Confusion Matrix for the second approach	29
5.4	Receiver Operating Characteristic (ROC) curve for the second ap- proach	30
5.5	Feature Importance	31
5.6	Biological Pathway Analysis	32
5.7	Protein-Protein Interaction (PPI) network highlighting interactions between identified proteins with a minimum score threshold of 0.3	35

List of Tables

4.1	Most significant genes as determined by SHAP value	23
-----	--	----

1. Introduction

1.1 Statement of Purpose

Schizophrenia (SCZ) and Bipolar Disorder (BD) are two common and severe psychiatric disorders that affect the well-being and functionality of individuals worldwide. These conditions make it difficult for afflicted individuals to participate in daily activities, resulting in social and economic burdens [46], [48]. According to data published by the World Health Organization (WHO), the estimated global prevalence of SCZ and BD was approximately 20 million [45] and 60 million [55] people, respectively in 2019. The prevalence and age distribution of SCZ and BD in 2021 are illustrated in Figures 1.1 and 1.2, highlighting the global burden of these psychiatric disorders.

BD is characterized by periodic episodes of mania, alternating with depression. Depressive episodes are marked by: **persistent sadness, irritability, or emptiness, often accompanied by a loss of interest in once enjoyable activities**. In contrast, during manic episodes, individuals may experience: **euphoria, irritability, or an increased sense of energy**. Additional symptoms can include: **rapid speech, racing thoughts, a reduced need for sleep, inattention** (e.g., an inability to keep track of tasks and activities with a high degree of destructibility), and **impulsive or high-risk behaviors**[45].

Similarly, SCZ involves impairments in perception and behavior, with individuals often experiencing: **delusions, hallucinations, disorganized thinking, and agitation**. Cognitive deficits, such as memory and attention problems, are also common in people living with SCZ[37]. Since SCZ and BD are associated with a higher risk of suicide and a lower life expectancy, both are presented as major contributors to global morbidity and mortality. Moreover, SCZ and BD constitute conditions whose diagnosis presents significant challenges, mainly due to their **overlapping symptomatology, the absence of definitive objective biomarkers, and the limitations inherent in traditional classification systems** [53].

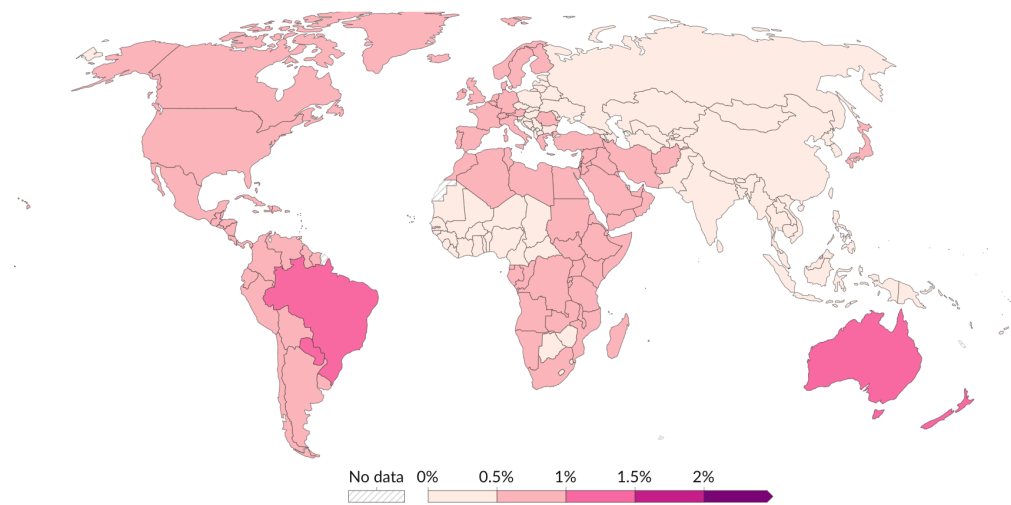
Previous studies have highlighted these issues, with a Danish cohort study that notably revealed a high degree of overlap between schizophrenia and bipolar disorder, leading to frequent misdiagnoses [32]. This finding underscores the need for a more nuanced diagnostic approach that considers the shared characteristics of these conditions. Other evidence from a study investigating the extent to which common genetic variations contribute to the risk of schizophrenia[13] corroborates the overlap in the genetic factors that appear in both SCH and BP. In the aforementioned study, two different analytical approaches were used: the Major Histocompatibility Complex (MHC) and the Polygenic Component. The MHC approach suggests that the MHC region, a part of the human genome, plays a role in the risk of SCH, where the Polygenic Component shows that many common genetic variants (thousands

of alleles) with small effects contribute to the risk of SHZ and also contribute to the risk of BD.

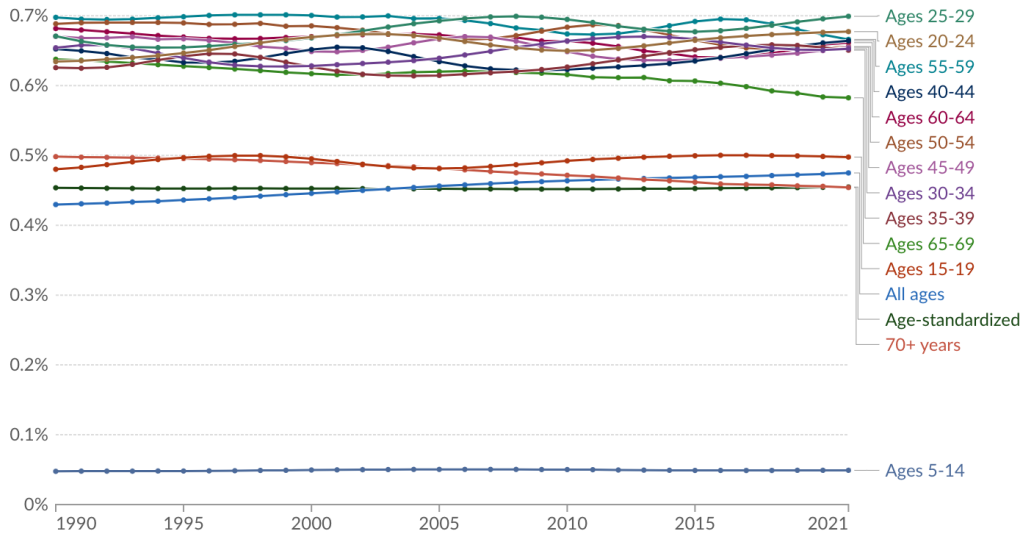
In light of the above, this study seeks to enhance the diagnostic strategy for SCZ and BD by integrating machine learning methods, which offer the potential to identify subtle patterns in patient data that traditional clinical interviews may overlook. A superficial definition of machine learning concludes the ability of machines to autonomously make decisions by analyzing patterns in the data without being explicitly programmed with specific instructions.

This approach complements conventional methods that rely primarily on clinical interviews and established diagnostic criteria. By leveraging the power of machine learning, more accurate and effective diagnostic tools are developed to better support the identification and treatment of these complex mental health conditions. In particular, as part of this research, a comprehensive data analysis pipeline was implemented. This includes:

1. the acquisition and preparation of an imbalanced data set that contains genetic information
2. the application of a machine learning model for multiclass classification and
3. the subsequent evaluation of the outcomes

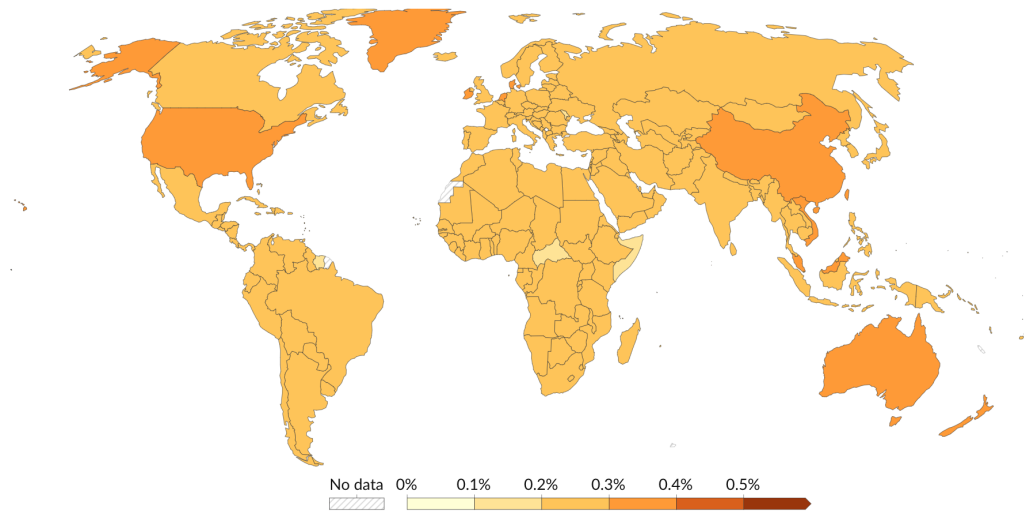


(a) Bipolar Disorder prevalence across the world, 2021. *Estimated share of males versus females who had bipolar disorder in the past year, whether or not they were diagnosed, based on representative surveys, medical data and statistical modelling.*

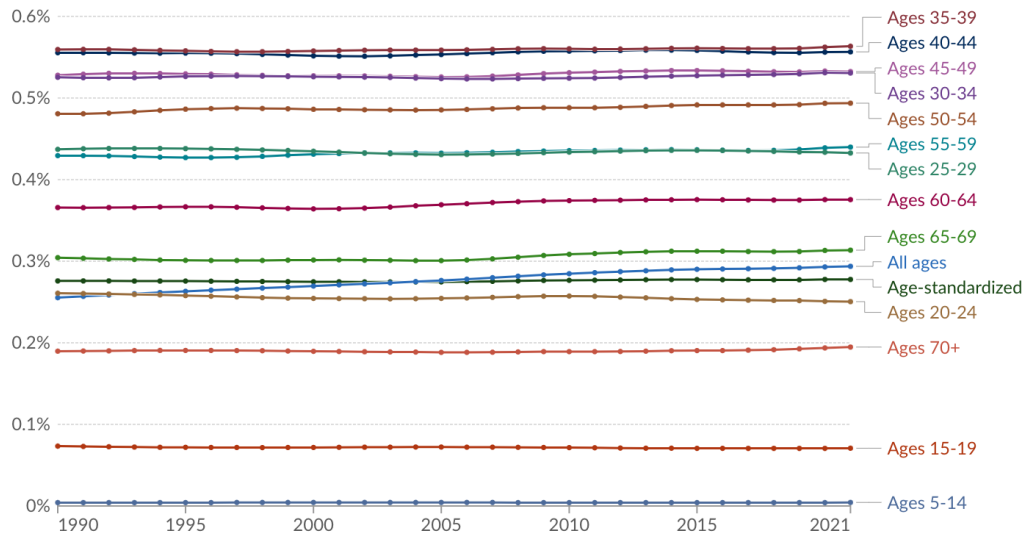


(b) Bipolar Disorder prevalence by age, 2021. *Estimated share of people who have bipolar disorder, whether or not they are diagnosed, based on representative surveys, medical data and statistical modelling.*

Figure 1.1: Bipolar Disorder prevalence, 2021[24].



(a) Schizophrenia prevalence across the world, 2021. *Estimated share of people who had schizophrenia in the past year, whether or not they were diagnosed, based on representative surveys, medical data and statistical modelling.*



(b) Schizophrenia prevalence by age, 2021. *Estimated share of males versus females who have schizophrenia, whether or not they are diagnosed, based on representative surveys, medical data and statistical modelling.*

Figure 1.2: Schizophrenia prevalence, 2021[25].

1.2 Objectives

The primary objectives of this research are:

- To train a machine learning model on genetic data to classify SCZ and BD.
- To identify genes that significantly influence the recognition of each psychiatric condition.
- To understand which biological functions or pathways are affected by the identified set of genes.

The remainder of this thesis is organized as follows.

- **Chapter 2**, presents the pivotal concepts and techniques.
- **Chapter 3**, presents an overview of existing applications and the importance of ML in the area of mental illnesses, as well as the research gaps that the present inquiry addresses.
- **Chapter 4**, presents the methodology utilized in this research.
- **Chapter 5**, presents the the outcome of the deployed ML model in comparison with the outcome reported by other investigations.
- **Chapter 6**, presents a summary of the key findings and contributions of this scientific investigation.

2. Theoretical Background

2.1 Machine Learning

Machine Learning, often abbreviated as ML, is a subcategory of artificial intelligence, widely known as AI. ML is a powerful tool that focuses on developing computer algorithms and statistical models to perform significant problem-solving tasks. At its core, ML enables electronic systems to operate autonomously, learn patterns and connections within large volumes of data, and make predictions. In other words, rather than relying on pre-programmed rules, ML algorithms improve their performance over time through exposure to data.

2.1.1 Key Types of Machine Learning

There are several types of ML, each with unique characteristics and applications. Key types of ML include Supervised Learning, Unsupervised Learning, Reinforcement Learning, Neural Networks, Ensemble Learning, and Instance-Based Learning (a comprehensive classification is demonstrated in figure 2.1).

This study employs a supervised machine learning approach to address a multiclass classification problem on imbalanced data; thus, the theoretical background primarily focuses on the key concepts and methodologies relevant to supervised learning. These include classification algorithms, ensemble methods, and techniques for handling class imbalances.

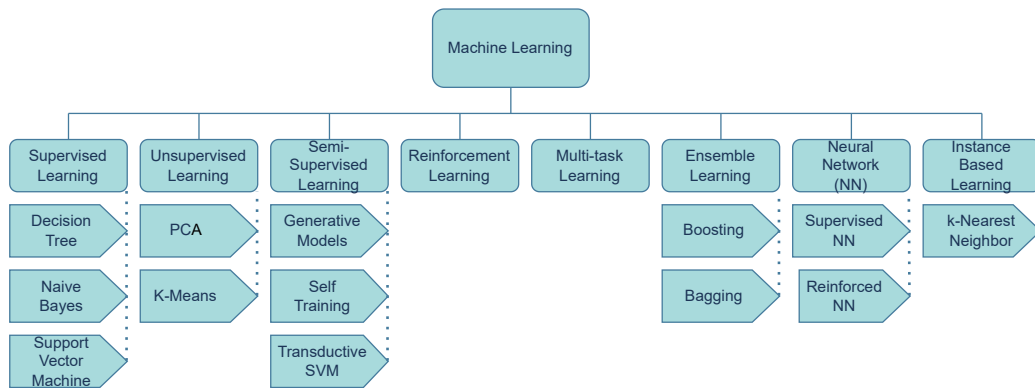


Figure 2.1: Key types of Machine Learning and algorithms[5]

An Overview of Supervised and Ensemble Learning

Supervised Learning: involves training a model using a labelled dataset that includes input features, usually presented as "X", and corresponding output labels,

usually presented as "y". **The purpose is to use and train an algorithm to map inputs to their corresponding outputs to make predictions on unseen data.** Supervised learning comprised two divisions, **Classification** and **Regression**.

1. Classification deals with categorical data that are used to solve categorical problems. The most well-known classification types are **Binary Classification** and **Multiclass Classification**. In Binary Classification, the goal is to classify the input into two categories (e.g. predict whether a tumor is benign or malignant). In contrast, in Multiclass Classification, the input is classified into one of several classes or categories (e.g., specify an animal's type among several other species). Here are some Traditional Machine Classifiers: Logistic Regression, Support Vector Machine (SVM), Random Forest, Decision Tree, K-Nearest Neighbors (k-NN) and Naive Bayes.
2. On the other hand, regression is a statistical method used to understand the relationship between **independent variables**, also known as **predictors**, and **dependent variables**, also known as **criterion**, aiming to predict the results in previously unseen data. The Analysis in Regression is conducted by observing how the dependent variable changes with regard to the independent variable. A typical example in medicine is predicting a person's blood pressure based on features (independent variables) such as age, weight, diet, and exercise frequency.

The three main uses of Regression Analysis are summarized in the following excerpts.

- Determining the Strength of Predictors: Determine which independent variable has the greatest impact on the independent variable.
- Forecasting an Effect: Predict the value of the dependent variable based on the value of the independent variable.
- Trend Forecasting: Understand patterns or trends in the data and make predictions based on future scenarios.

Here are some classification algorithms: linear regression, polynomial regression, and decision tree regression.

Ensemble Learning: refers to a methodology within the broader category of Supervised Learning, commonly used in both Classification and Regression tasks. By definition, an ensemble denotes a unit or group of complementary components that contribute to a unified outcome.

In machine learning, Ensemble Learning combines the strengths of multiple models to improve the accuracy of predictions. One of the techniques that Ensemble Learning includes is a key approach known as Boosting. Boosting is considered an advanced ensemble technique in which base models are trained sequentially,

with each subsequent model addressing the weaknesses of the previous ones. The final prediction is a weighted average of the individual models' predictions with higher weights given to more accurate models. Intuitively, ensemble learning resembles a group of people trying to guess the answer to a question. Initially, the first person predicts the answer to the question (**base model**). Based on this, the next person is trained to make predictions focusing on correcting the mistakes of the previous (**sequential training**). The final prediction is the weighted average of all the projections made by the group of people(**weighted summary**).

Boosting Algorithm under Loss Function L

Inputs: Joint distribution Q on $\mathcal{Z} = \mathcal{X} \times \{-1, +1\}$ and initial predictor $H^{(0)} \in \mathcal{S}[\mathcal{H}]$. In practice, Q is the empirical probability of training samples.

Loop For $m = 1, \dots, M$

1. Find a hypothesis $h^{(m)} \in \mathcal{H}$ such that

$$h^{(m)} = \arg \min_{h \in \mathcal{H}} \int_{\mathcal{Z}} Q(dz) L'(-yH^{(m-1)}(x)) I(y \neq h(x)).$$

The minimization does not need to be exact.

2. Find a coefficient $\alpha^{(m)} \in \mathbb{R}$ such that

$$\alpha^{(m)} = \arg \min R_L(Q, H^{(m-1)} + \alpha h^{(m)}).$$

Line search or Newton methods can be applied to solve the one-dimensional optimization problem.

3. Update the predictor:

$$H^{(m)} = H^{(m-1)} + \alpha^{(m)} h^{(m)}.$$

Output: $H^{(M)}$ as an estimated predictor.

Figure 2.2: Boosting Pseudo code[5]

2.1.2 Gradient Boosting Decision Trees

Decision Tree

A Decision Tree (DT) is a tree-like decision support model consisting of a series of choices and their possible outcomes. As seen above, DT is a non-parametric super-

vised learning method for classification and regression tasks. The key components of DTs are the following:

1. **Root Node:** The starting point of the DT.
2. **Internal Nodes (Decision Nodes):** Nodes that succeed the root node.
3. **Leaf Nodes (Terminal Nodes):** All possible outcomes based on the data.

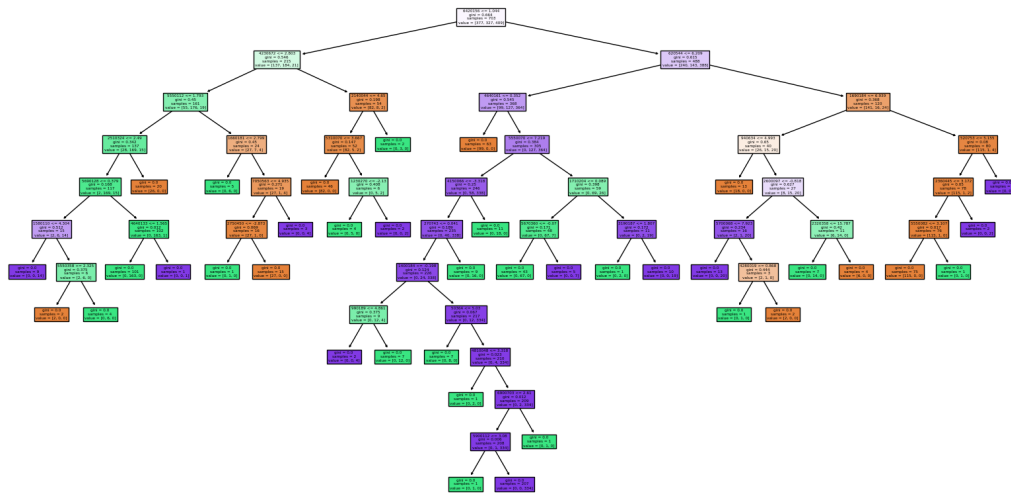


Figure 2.3: The tree structure generated from the dataset used in this study, following training with a Decision Tree Classifier.

Gradient Boosting Decision Trees (GBDT) is a specific implementation of Boosting. The model works the same way as ensemble learning, combining weak learners to predict the remaining inaccuracies of the previous one. In GBDT, decision trees are used as the basic learners. Each tree is connected, following a sequential arrangement, and attempts/strives/makes a strong effort to correct the mistakes made by the preceding tree. This procedure begins with the attempt of a weak learner to predict the target outcome. Following this event, the model performance is evaluated, and the errors, marked as residuals, are identified. Then, a subsequent weak learner is introduced to the framework to predict the residuals from the initial model. This process is iterative, with each new weak learner focusing on minimizing the residuals from the previous step. Over successive iterations, every new learner fits into the residuals of the previous step, resulting in an increasingly accurate model. Ultimately, the final model integrates the contributions of all weak learners to form a robust predictive model. A concise description of the operational framework of the Gradient Boosted Decision Tree (GBDT) technique is presented in 2.4.

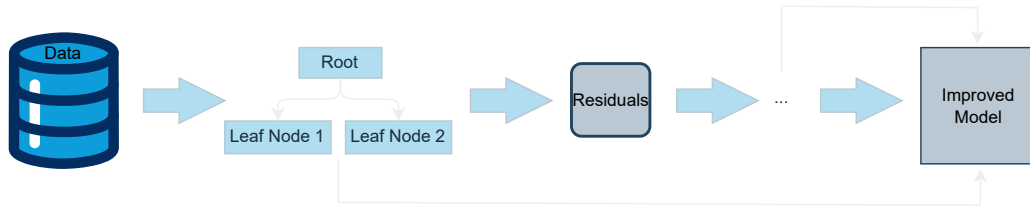


Figure 2.4: Operational framework of the Gradient Boosted Decision Trees (GBDT)

2.1.3 XGBoost: A Boosting Algorithm

XGBoost, or eXtreme Gradient Boosting [10], is a powerful and efficient library implementation of Gradient Boosting designed for scalability and high performance. Known for its ability to handle large datasets and high-dimensional feature spaces, XGBoost incorporates advanced techniques such as regularization, parallelized tree construction, and sparsity awareness to deliver state-of-the-art results across diverse machine learning challenges [11]. XGBoost iteratively builds decision trees, creating multiple trees that work together to make predictions. It optimizes an objective function, a mathematical formula that helps the model learn from the data. Its combination of computational efficiency and predictive accuracy makes it a popular choice in applications such as genomics, where datasets are often characterized by high dimensionality and imbalance. Lastly, it is particularly good at handling structured data, like numbers and categories.

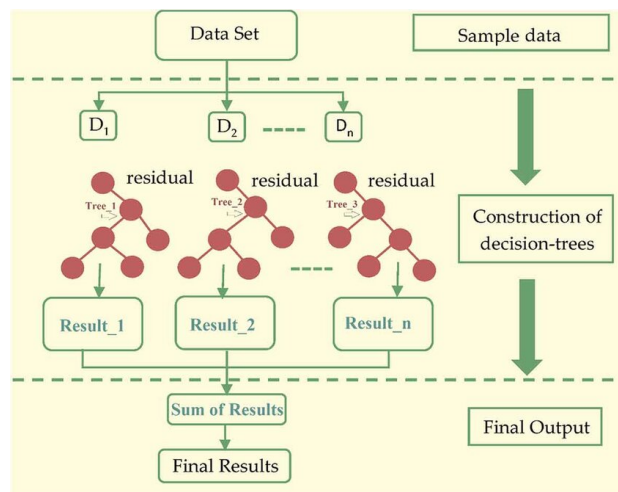


Figure 2.5: General architecture of XGBoost algorithm [1].

2.2 Bipolar Disorder and Schizophrenia

SCZ significantly affects people's thoughts, feelings, and behavior. The symptoms are divided into three categories: **Positive**, **Negative**, and **Cognitive**. **Positive symptoms** include hallucinations, delusions, disorganized speech, and unusual movements. **Negative symptoms** include anhedonia (lack of pleasure), alogia (speech problem), lack of emotions, and isolation. **Cognitive symptoms** include cognitive impairment, distraction, and disorganization.

BD is identified by mood disturbances. The symptoms are divided into six categories, **Manic Episodes**, **Hypomanic Episodes**, **Depressive Episodes**, **Mixed Episodes**, **Psychotic Symptoms**, and **Cyclothymic Disorder**. **Manic episodes** include excessive happiness, high energy, sleep deprivation, a "flight of ideas", and destructibility. **Hypomanic episodes** are similar to manic episodes but milder. **Depressive episodes** include nihilistic ideation (distorted self-perception), suicidal ideation, sleep disturbances, change in appetite, and cognitive dysfunction. **Mixed episodes** involve a mixture of symptoms of mania and depression. **Psychotic symptoms** occur during severe episodes of mania or depression and include hallucinations, delusions, and paranoia. **Cyclothymic disorders** include frequent mood swings.

2.2.1 Anatomical, Biological, Biochemical, and Genetic Perspectives on BP and SZ

BD and SZ are complex psychiatric disorders with shared symptoms but also distinct characteristics. To expand our understanding of these disorders, an interdisciplinary approach that encompasses anatomical, biological, biochemical, and genetic perspectives is necessary.

Anatomical Perspective

Study [16] shows that both schizophrenia (SZ) and bipolar disorder (BD) involve changes in brain structure. In SZ, brain parts that control thinking and memory, such as the prefrontal cortex and hippocampus, shrink more significantly than in BD. People with BD also show changes in brain areas related to emotion, like the amygdala. Both disorders have disruptions in brain connections (white matter), but these are more severe in SZ. These differences help researchers understand how the brain works in each disorder and guide the diagnosis.

Biological Perspective

SZ and BD involve problems with brain chemicals like dopamine, glutamate, and serotonin, which influence mood, thinking, and behaviour. In SZ, dopamine imbalances cause symptoms like hallucinations, while in BD, they contribute to mood

swings. Both disorders also show signs of inflammation, meaning the immune system may play a role. Understanding these shared and unique biological changes helps find better treatments for each condition [22],[36].

Biochemical Perspective

Both SZ and BD show signs of stress on the body at a cellular level, like damage caused by unstable molecules (oxidative stress) and problems with energy production in cells (mitochondrial dysfunction). Research has shown that mitochondrial dysfunction, particularly *complex-I* impairment, and increased oxidative damage in the prefrontal cortex are implicated in the pathophysiology of both disorders [3]. Stress hormones are also overactive in both disorders, affecting the brain's ability to manage stress. These changes might explain symptoms like memory problems and mood instability. Treatments often target these stress-related changes to help stabilize symptoms.

Genetic Perspective

SZ and BD familial disorders, meaning that they run in families, indicating a significant genetic contribution to their etiology. The study by Lichtenstein et al. [33] addresses the genetic and familial contributions to these disorders, showing how genetic influences and environmental factors shape their development. It discusses the shared genetic factors between SZ and BD, supporting the observation that some genetic factors are shared while others are unique. For example, genes related to brain communication and immune system regulation are related to SZ, while genes that affect mood and sleep patterns are more related to BD. Additionally, stressful life events can also trigger these disorders by altering gene functionality. Understanding these genetic and environmental influences helps identify at-risk individuals and develop personalized treatments.

In conclusion, SZ and BD share similarities, like changes in brain structure and inflammation, but also have clear differences. SZ is characterized by more widespread brain changes and problems with thinking, whereas BD is more focused on mood regulation. While an integrative approach that combines brain imaging, genetic, and biological studies enhances the accuracy of diagnosis and treatment, these methods can be costly and time consuming. In contrast, genetic-based approaches offer a faster and more cost-effective alternative; however, distinguishing between SZ and BD solely using genetic data remains a significant challenge due to the complex and overlapping genetic underpinnings of these disorders. Addressing this challenge requires further advancements in genetic profiling and computational models to improve classification accuracy.

3. Literature Review

This chapter presents an overview of the relevant literature on methods applied for the diagnosis of mental health diseases focusing on SCZ and BD. It highlights inefficiencies in traditional diagnostic methods and relevant ML techniques that aim to tackle them.

3.1 Traditional Diagnostic Methods

Traditional diagnostic methods for SCZ and BD primarily involve the combination of a) **Clinical Interviews and Mental Status Examinations**, b) **Patient History and Collateral Information**, and c) the application of standardized criteria based on diagnostic manuals such as the **Diagnostic and Statistical Manual of Mental Disorders (DSM)** and the **International Classification of Diseases (ICD)**. These manuals not only serve as a guide for mental health professionals to identify and categorize mental health conditions but also facilitate the establishment of a common ground for clear and consistent communication between them.

Elaborating on these specific tools and criteria, healthcare providers/clinicians conduct comprehensive interviews to evaluate the patient's mental capacity, including cognition (focus, spatial awareness, memory, reasoning, and judgment), mood, and affect behaviour and perception. Observations are made based on specific insights, including the individual's

- **personal hygiene and physical appearance,**
- **levels of distress, cooperation, and discomfort,**
- **abnormal motor manifestations**, such as hyperactivity, tremors, tics, and rigidity
- **frequency and fluency** in verbal communication
- **presence or absence of emotional expression**
- **cognitive organization and thought process**

The DSM and ICD outline specific criteria characterized for SCZ and BD.

3.2 Diagnostic Methods Using Machine Learning

Machine learning methods offer the potential to identify subtle patterns in patient data that traditional clinical interviews may overlook. By analysing vast amounts

of information from various sources, such as blood samples, medical images, and so on, ML algorithms can identify complex relationships and correlations that may not be apparent through manual examination alone. This enables healthcare professionals to gain a deeper understanding of a patient's condition, leading to more accurate diagnoses and personalized treatment plans. Nevertheless, distinguishing between schizophrenia (SCZ) and bipolar disorder (BD) remains a formidable challenge due to significant genetic and molecular overlaps between the two conditions. Consequently, current diagnostic procedures can take an average of 10 to 15 years [34], [15], delaying timely interventions and treatments [21]. Given this complexity, machine learning (ML) techniques are uniquely positioned to enhance differentiation between SCZ and BD, reducing diagnostic delays and improving patient outcomes.

In [2], the author analyzed data from individuals born between 1981 and 2005, focusing on psychiatric diagnoses and genetic information, with anonymized data approved by the Danish Data Protection Agency. Specifically, genetic data were scaled between 0 and 1 with missing values imputed and categorical data encoded using feature embedding. All post-diagnosis information was removed to emulate the clinical data available at the time of the diagnosis, and only prior diagnostic data were kept. The study aimed to create a model that can predict mental health diagnoses and how severe they will be over time. For the predictive modelling, a feed-forward Neural Network (NN) architecture was implemented in PyTorch and evaluated using a stratified 3-fold cross-validation scheme. Based on the aforementioned architecture, two models were developed: a) A single multiclass model designed to predict multiple disorders simultaneously. b) Separate binary prediction models for each disorder, focusing on distinguishing between a specific disorder and all other disorders. The model's performance was evaluated using several metrics including the Area Under the Curve (AUC) and Matthews Correlation Coefficient (MCC). To address the class imbalance, accuracy was further assessed using the lift, defined as the ratio between actual and expected accuracy under random predictions. The results from the prediction of severe mental disorders in a Background Population are as follows:

- **Model 1:** The binary prediction model achieved an AUC of 0.72, an accuracy of 66% (representing a 1.2-fold lift from a by-chance accuracy of 56%), and an MCC of 0.27.
- **Model 2:** The multiclass model predicted specific mental disorder diagnoses with an AUC of 0.81 and an accuracy (equal to the positive predictive value for multiclass models) of 44% (a 1.9-fold lift from by-chance accuracy of 23%).
- **Model 3:** The multidiagnostic model, when restricted to clinical cases only, yielded an overall AUC of 0.82, MCC of 0.36, and an accuracy of 53% (2-fold lift from by-chance accuracy of 27%).

The study [2] shares similarities with the presented work, as both use machine learning with clinical data, focusing on metrics like AUC under class imbalance. While [2] applied a feed-forward neural network, this study used XGBoost, achieving higher accuracy (92.87% in the imbalanced dataset and 0.90% in the balanced dataset vs. 66% and 53%). Limitations of [2] include insufficient methods for handling class imbalance.

In [57], researchers explored the application of machine learning algorithms in differentiating psychiatric disorders such as schizophrenia (SCZ), bipolar disorder (BP), and major depressive disorder (MDD). They constructed a dataset comprising 268 samples: 67 SCZ, 40 BP, 57 MDD, and 104 healthy controls (CTRL). The study integrated gene expression data from three transcriptomic datasets - GSE92538, Stanley AltarC, and GSE53987 - focusing on the dorsolateral and anterior prefrontal cortex (DLPFC), specifically Brodmann areas (BA) 9, 10, and 46. These regions are involved in cognitive functions such as memory retrieval, reasoning, task flexibility, problem-solving, planning, execution, working memory, processing emotional stimuli, inferential reasoning, decision-making, and numerical processing.

To mitigate systematic differences between the datasets, batch effect correction was applied. The researchers applied PLS-DA, a high-dimensional data-specific feature selection technique. Differentially Expressed Genes (DEGs) were filtered based on a VIP score above 2 and validated through literature reviews, protein-protein interaction network analysis, and gene ontology term analysis to ensure that only the most relevant biomarkers contribute to classification.

The classification model was constructed using *Support Vector Machines (SVM)*. Since SVM inherently handles only binary classification, the researchers employed the *One-vs-Rest* approach to address this limitation. This method involves training multiple SVM models, each designed to distinguish one specific class from all others. To assess the model's performance, they used *five-fold cross-validation*, which resulted in an *average area under the curve (AUC) value* of 0.94 for the combined dataset. When tested in an independent dataset (*GSE127711 and GSE38484*), where gene expression was measured in blood samples rather than brain tissue, the model achieved an AUC of 0.71, indicating reduced performance due to differences in tissue types (blood vs. brain).

Even though these results are promising, it must be pointed out that the *One-vs-Rest* approach can lead to overfitting, where the model is too specialized in the training data and becomes less competent on unseen data. Furthermore, mixing gene expression data from different tissue types in the training and independent sets complicates the assessment of models, leading to potential biases and an unfair approximation of their real performance.

In [51], the researchers investigate the ability of ML to identify individuals at high risk for schizophrenia (SCZ) by analyzing *Whole Exome Sequencing (WES)* data. The data collected via the database of *genotypes and phenotypes (dbGaP)* comprised 2,545 individuals with SCZ and 2,545 unaffected individuals. Genetic

variants, including *single nucleotide variants (SNVs)* and *small insertions and deletions (indels)*, were annotated using *ANNOVAR* with the reference genome *hg19/GRCh37*. Rare predicted functional variants with a minor allele frequency (MAF) < 1% and genotype quality less than 90 were selected for analysis.

The study employed a supervised ML approach, using *XGBoost* with regularization to mitigate overfitting. Data were split into 70% for training and 30% for testing. To enhance feature selection, the authors filtered out features with *Pearson correlation > 90%* and *standardized feature values to have a mean of 0* and a *standard deviation of 1*. From an initial 17,138 features, only 1,155 relevant features were retained through L1-regularized logistic regression, random forests, and *XGBoost*'s implicit feature selection.

The performance of *XGBoost* was assessed using accuracy, specificity, sensitivity, and the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve. The results demonstrated a high prediction accuracy of 85.7%, specificity of 86.6%, sensitivity of 84.9%, and an AUC of 0.95. Feature importance analysis identified the top 50 predictive genes, which were further analyzed using bioinformatics tools such as DAVID. The most overrepresented pathways included the **MAPK signaling pathway (hsa04010)**, **JAK-STAT signaling pathway (hsa04630)**, **cGMP-PKG signaling pathway (hsa04022)**, and **calcium ion signaling pathway (hsa04020)**.

One limitation of [51] is the absence of an in-depth investigation into the functional contribution of the identified important genes to schizophrenia.

Gaps and Challenges

To the best of our knowledge, no study has addressed a multiclass classification problem to distinguish SCZ and BD using genetic data while leveraging boosting algorithms that work well in imbalanced gene datasets commonly available for biomedical studies. Most previous studies have framed SCZ and BD classification as a binary problem, typically comparing each disorder to controls, or have employed the *One-vs-Rest* approach. Commonly used methods include Meta-analytic Cognitive Priors, Logistic Regression (LR), Support Vector Machines (SVM), and Random Forest (RF).

Moreover, given the effects of neurodevelopmental psychiatric disorders on brain structure, most studies rely on neuroimaging data from MRI, fMRI, or sMRI screenings. However, obtaining such data requires expensive and time-consuming processes. This study addresses these gaps by developing an ML framework that distinguishes SCZ and BD from controls in a multiclass setting while leveraging boosting algorithms to enhance classification performance. Furthermore, it integrates gene importance analysis with pathway enrichment to provide insights into the biological underpinnings of these psychiatric conditions. Finally, this work advances the methodological landscape of psychiatric genetics and the biological interpretation of genetic markers relevant to SCZ and BD through GO enrichment

analysis and PPI on genes the model deems important for the classification process.

An additional benefit of the above analysis lies in its multi-classification approach. Given the biological and clinical complexity of psychiatric disorders such as SCZ and BD, which share genetic risk factors and symptomatology, this approach helps to identify **disease-specific markers** rather than a general set of vulnerability markers. Unlike binary classification, which strictly separates disorders, a multiclass approach offers a more detailed view of the psychiatric spectrum, highlighting both common and unique molecular patterns across disorders.

4. Methodology

This chapter outlines a comprehensive overview of the data analysis pipeline employed in this study, from sourcing and preparing the data to applying machine learning models and evaluating their performance. Particular attention is given to the approach for handling imbalanced data, which is a critical concern in classification tasks.

4.1 Dataset Description

The gene expression dataset used in the present study was obtained from the ArrayExpress repository under the accession number E-MTAB-8018 [8]. The data were sourced from a study published by Chagnon et al. (2019) investigating the genetic underpinnings of SCZ and BD through a multimodal genomic approach. The dataset comprised, **RNA expression**, **Single Nucleotide Polymorphism (SNP) genotyping** - a way to identify small variations in the DNA sequence of a gene and **DNA methylation measurements** from 498 human subjects. These subjects include individuals diagnosed with SZ, BD, and non-affected relatives(NAR).

4.1.1 Data Collection and Phenotype Definition

Chagnon et al. (2019) collected samples from a subset of 29 out of the original 48 kindreds in their study. These kindreds included both SZ and BD patients, as well as their NARs. In total, out of the 156 affected subjects, 58 were diagnosed with SCZ, 13 with schizoaffective disorder (SAD), and 85 with BD, while 342 non-affected subjects were also included.

Each subject was diagnosed using a best-estimate lifetime DSM-IV diagnosis. This diagnosis involves a thorough review of medical records, family interviews, and structured interviews conducted by four independent research diagnosticians.

For the analysis in this current study, the gene expression and SNP genotyping data were utilised to create a machine learning model with the goal of predicting disease phenotype based on genetic factors. The dataset was derived from peripheral blood samples chosen for their accessibility and ability to represent systemic biological processes. This approach provides a snapshot of the body's overall health rather than focusing on a specific area. Specifically, RNA was extracted from lymphocytes, a critical type of immune cell in the blood.

4.1.2 Dataset Details

Expression values have been normalized and are accessible via File Transfer Protocol (FTP). The provided demographic information includes a gender breakdown (male/female) for each cohort, aged 30 to 83 years. Data were generated through micro-array analysis using biotin labelling performed on the A-GEOD-10558 platform.

All experiments were constructed according to a standardized protocol, *P-MTAB-86516*, to avoid inconsistencies and ensure sample reliability. This dataset provides invaluable insights into the transcriptomic differences linked to SZ and BD, contributing to the understanding of the molecular mechanisms underlying these disorders.

4.1.3 Data Pre-processing Methods

To identify genes that are consistently expressed across a significant portion of the study population, the study team employed a three-step process to analyze RNA expression data obtained from Illumina chips, preparing them for downstream analysis. The methodology involves the following steps:

1. **Data normalization:** Probe measurements were adjusted to remove background noise and ensure the data was on a comparable scale across all chips (Background correction and quantile normalization).
2. **Data Transformation:** The normalized data were then converted to a logarithmic scale (\log_2) to quantify gene expression levels accurately. The transformed data were subsequently used to determine the expression levels of each gene, represented by individual probes.
3. **Data Filtering:** Probes were evaluated using the Illumina detection test, and only those with a p-value < 0.05 were considered expressed. Additionally, probes expressed in at least 75% of the subjects from one or more families (with a minimum of four members measured per family) were retained for further analysis.

The aforementioned methods ensure that the data is reliable and adequately prepared for further analysis, such as training machine learning models.

4.2 Machine Learning Model

In this research, the ML model employed is the XGBoost Classifier, which was configured with the following parameters:

- **Objective:** *multi:softmax*. The goal of the model is to perform multi-class classification, which assigns each instance to one of the predefined classes.
- **Number of Classes:** 3. The model is configured to handle three distinct classes in the dataset.
- **Tree Construction Method:** *hist*. The model uses a histogram-based approach to build the decision trees. This method is computationally efficient and accelerates the tree-building process by grouping continuous data into discrete bins.
- **Early Stopping Rounds:** 10. Early stopping is implemented as a regularization technique to monitor the model's performance on a validation set. Specifically, the training process is halted if the evaluation metric doesn't improve after 10 consecutive rounds. This prevents overfitting by terminating training when further iterations have little impact. Conversely, if the log loss continues to improve, the model will continue training until the specified number of maximum rounds is reached.
- **Evaluation Metric:** *mlogloss*. The multi-class logarithmic loss is used as the evaluation metric to assess the model's performance in predicting probabilities, with lower values (closer to zero) reflecting better predictions.

These parameters ensure that the XGBoost Classifier is optimized for multi-class classification while computational efficiency is optimized and overfitting is minimized.

4.2.1 Training Methodology

The model was trained in two distinct approaches. The first approach employed 5-fold Stratified K-Fold Cross-Validation to ensure a balanced representation of each class in the training and test subsets. The training pipeline included:

- **Train-Test Splitting:** Each fold was further split into training and validation sets to enable early stopping.
- **Training with Early Stopping:** The model monitored validation performance to stop training once no significant improvement was observed for 10 consecutive rounds.
- **Evaluation:** Predictions were made on the test subset of each fold for subsequent metrics calculation.

The second approach utilised a manually balanced dataset for the training, specifically, the process involved:

- **Manual Train-Test Splitting:** The dataset was split into training and test sets for the model's training and evaluation.
- **Training with Early Stopping:** Similar to the first approach, the model used validation performance to halt training after 10 epochs without notable improvement.
- **Evaluation:** Predictions were made on the test subset to assess the model's performance.

4.2.2 Tools, Frameworks, and Setup

The development environment leveraged the following tools and frameworks:

- **Python:** Primary programming language.
- **XGBoost:** The machine learning library for gradient boosting.
- **scikit-learn:** Utilized for metrics calculation, preprocessing (e.g., label binarization), and splitting datasets.
- **Matplotlib and Seaborn:** For visualization, including confusion matrices and ROC curves.
- **NumPy:** For array manipulation and numerical computations.

4.3 Evaluation Metrics

To comprehensively evaluate the model's performance, the following metrics were used:

- **Accuracy:** to measure the overall correctness of the classification model.

$$\text{Accuracy} = \frac{\text{True Positives (TP)} + \text{True Negatives (TN)}}{\text{Total Number of Instances}}$$

Despite this measurement's effectiveness on balanced datasets, it can be misleading for imbalanced ones, as it favours the majority class. To better judge the performance of the model under imbalanced data the following metrics are also utilised.

- **Precision:** to measure the percentage of instances correctly predicted as positive. This measurement is also known as Positive Predictive Value (PPV).

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

- **Recall:** to measure how many positive instances were correctly identified. It is also called Sensitivity or True Positive Rate (TPR).

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

- **F1 Score:** to evaluate classification models' performance, particularly in imbalanced class distributions. It is the harmonic mean of precision and recall.

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Confusion Matrix:** a simple table that clearly shows how well the model classifies each class in the multi-class classification problem. The model evaluation matrix categorizes the predictions into four types: True Positives (TP) represent the model's ability to correctly identify the positive outcome, True Negatives (TN) represent the model's ability to correctly identify the negative outcome, False Positives (FP) represent the model's ability to incorrectly identify a positive outcome (Type I error), and False Negatives (FN) represent the model's ability to incorrectly identifies a negative outcome (Type II error).
- **Receiver Operating Characteristic - Area Under the Curve (ROC-AUC):** The Receiver Operating Characteristic (ROC) curve provides a visual representation of a model's performance across various classification thresholds. In this context, a threshold determines the decision boundary for classifying instances as positive or negative. The Area Under the Curve (AUC) quantifies the model's ability to rank positive instances (e.g., individuals with the disease) higher than negative instances (e.g., healthy individuals). The ROC-AUC curve is a fundamental tool for evaluating the performance of binary classification models. For **multiclass classification problems**, the ROC curve can be extended using the **One-vs-Rest (OvR) approach**. In this method, a separate **ROC curve is computed for each class**, treating it as the positive class while considering all other classes as negative. The overall performance can then be summarized using either the **macro-average AUC**, or the **weighted-average AUC**.
- **Micro-Average AUC:** aggregates the contributions of all classes by globally summing the true positives, false positives, and false negatives before

Gene Names
LRRC32, LOC400707, C12ORF54, CCDC109A, ZNF91, MIR641, LOC100134282, SLC48A1, PEX5, LOC651115, LOC100128269, PDIA6, LSM3, FLJ32310, GAGE2C, LOC23117, TRIM61, LOC100129673, HS.576698, GRIN2A, LOC645381, LOC643246, LOC647666, LOC651398, PCAF, KRT10, FNBP4, HS.582914, RALA, C15ORF39, LOC728937, MED18, CPO, ABHD12, JARID1D, PCD-HGA9, CCNB1IP1, LOC730004, LOC651101, LOC440243, TCEA2, RPL14, DDX19B, FLJ10781, FLJ12355, MEG8, CASP12, LOC728255, UMODL1, SP8, LOC100134498, TMEM59L, NPY, STK35, BCL2L1, LUZP4, RAB22A, OR52I1, CDK5R1, CAPN3, UGT2B7, COX7A2, LOC150763, THAP10, POLR2J2, OTOS, HS.549784, CCDC23, CHRM4, ASTL, LOC730294, LOC652051, LOC644891, ANKRD30A, GLP2R, FOXA1, NRM, MBNL1, SRP14, LOC282997, PYY, KIF25, LOC651936, EIF3G, LOC100132299, LOC731682, CENPA, MGC3771, ITGB4BP, LOC728601

Table 4.1: Most significant genes as determined by SHAP value

calculating the AUC. In essence, it gives more weight to the larger classes (if there are class imbalances) since it reflects the overall performance across the dataset. This makes it sensitive to class sizes, so larger classes contribute more to the overall score.

- **Macro-Average AUC:** is the arithmetic mean of the AUC scores calculated separately for each class. It treats all classes equally, regardless of size (i.e., class imbalance is not considered).

4.3.1

The data preparation process for training the machine learning model is employed through two distinct approaches, as outlined in the following sections.

First Approach: In the first approach, gene expression data, along with labels and cohort characteristics related to age and gender, were collected to build the data set. Features (X) and labels (y) were extracted from an imbalanced dataset. Given that XGBoost is inherently robust and capable of handling class imbalance, no additional balancing strategies (e.g. `scale_pos_weight` or class weights) were applied during this stage.

The dataset was partitioned into five folds using the *StratifiedKFold* method to ensure reliable evaluation and consistent class distribution across splits. This stratified approach preserved the proportion of each class in every fold, providing a balanced framework for cross-validation and mitigating potential biases in model's

assessment.

Predictions for each fold were stored to compute overall performance metrics and plot visualizations. Confusion matrices were plotted to analyze misclassification while ROC curves for individual classes and average performance were generated to assess separability.

Second Approach: In the second approach, gene expression data, labels and cohort characteristics were again used to construct the dataset. However, this time, the dataset was processed in a way that eliminates class imbalance. Using the *RandomUnderSample* technique, the larger classes (control and SCZ) were reduced in size to match the number of the minority class (BD). Initially, the dataset comprised, 391 control samples, 84 SCZ samples and 58 BD samples (533 samples in total). After undersampling, each class contained 58 samples, resulting in 174 total samples.

The dataset was manually partitioned into training and test sets. Out of the 174 total samples, 10 samples from each class were randomly selected to form the test set, resulting in a total of 30 test samples. The remaining 144 samples were used for training. The XGBoost model was trained and evaluated on this processed dataset.

To improve performance, feature selection was applied using SHAP (SHapley Additive exPlanations) values, identifying the 90 most significant genes. These genes are shown in table 4.1.

The model on the second approach was retrained using the dataset containing only these selected features. Predictions were evaluated using metrics such as accuracy, precision, recall, F1 score, AUC, and the confusion matrix to provide a comprehensive assessment of model performance.

5. Results

5.1 Overview of Model's Performance

The XGBoost classifier was trained and evaluated using a stratified 5-fold cross-validation approach, ensuring a robust assessment of its generalization performance. The results demonstrate the model's strong performance in distinguishing between the three classes of the target variable. The following sections detail the key evaluation metrics, including the confusion matrix, Receiver Operating Characteristic (ROC) curve analysis, and the classification metrics presented in section 4.3.

Results of the First Approach

The confusion matrix, as represented in Figure 5.1, provides a clear breakdown of the model's classification performance across the three classes. Class 0, 1 and 2 represent the control group, individuals diagnosed with BD, and individuals diagnosed with SCH, respectively. The model performs well in identifying instances of Class 0, with no misclassification. For Class 1, the model correctly classifies 79.8% of samples; however, a misclassification rate of 20.2% is observed, with these samples being incorrectly assigned to Class 2. Similarly, the model correctly identifies 63.8% of Class 2 samples, although 36.2% are misclassified as Class 1. These results indicate that while the model effectively distinguishes Class 0, it encounters greater challenges in differentiating between Classes 1 and 2.

The ROC-AUC curves, as represented in Figure 5.2, illustrate the trade-off between the true positive rate (TPR) and the false positive rate (FPR) for each class. The results are interpreted as follows: an AUC of 1 indicates perfect classification, while an AUC of 0.5 indicates a model that performs no better than random guessing. The results gathered from the model reflect its ability to distinguish effectively between all three classes. Notably, the perfect AUC for Class 0 indicates the model's exceptional precision for this class. Where the AUC values gained for Classes 1 and 2 are also satisfactory.

A micro-average AUC of 0.97 shows that the model effectively distinguishes between true and false positives across all instances in the dataset (irrespective of class).

A macro-average AUC of 0.97 indicates that the model performs consistently well across all classes, even for the smaller or less frequent ones. This suggests that the model's ability to discriminate between positive and negative instances is strong across the three classes in the presence of class imbalance.

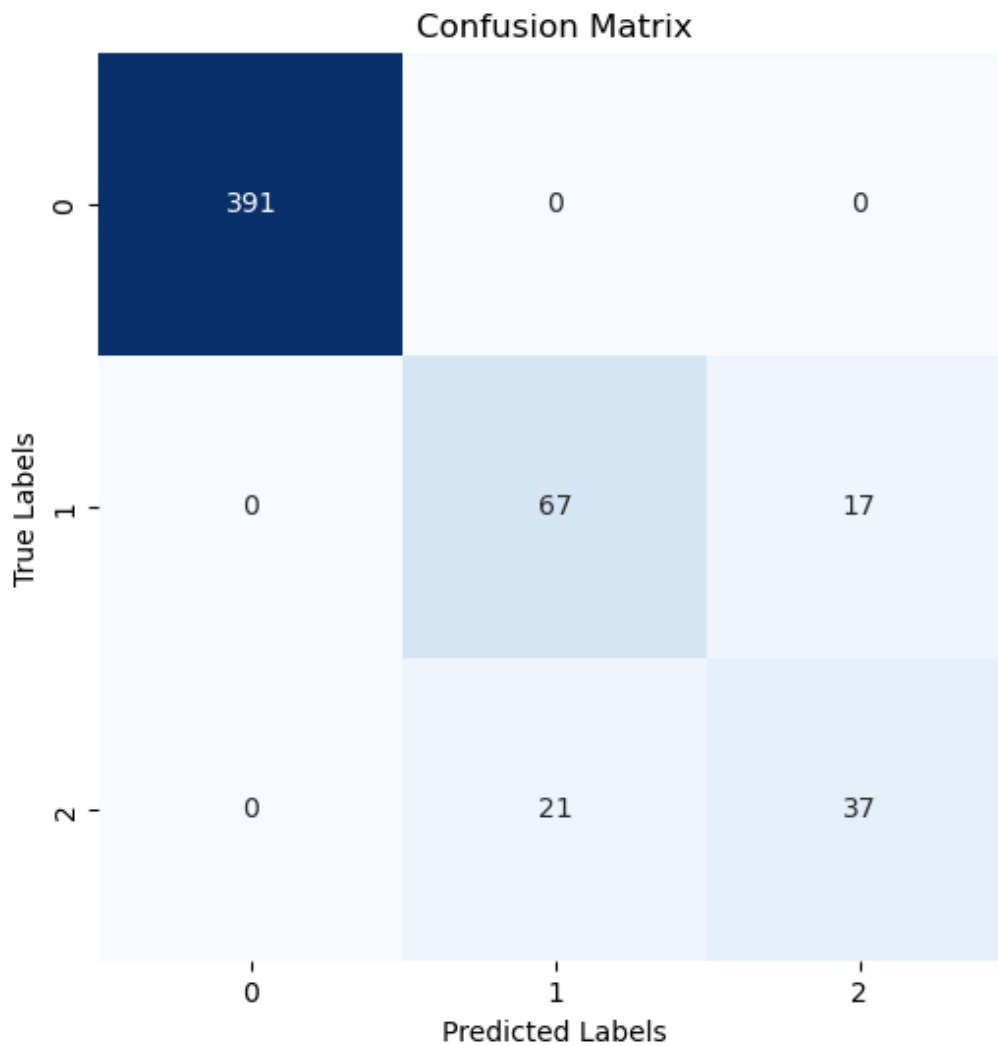


Figure 5.1: Confusion Matrix for the first approach

Confusion matrix for the XGBoost model on an imbalanced dataset, illustrating classification performance. The x- and y-axes represent the control (0), schizophrenia (1), and bipolar disorder (2) groups. Diagonal elements, from the top left to the bottom right, indicate correctly classified instances, while off-diagonal elements represent misclassifications.

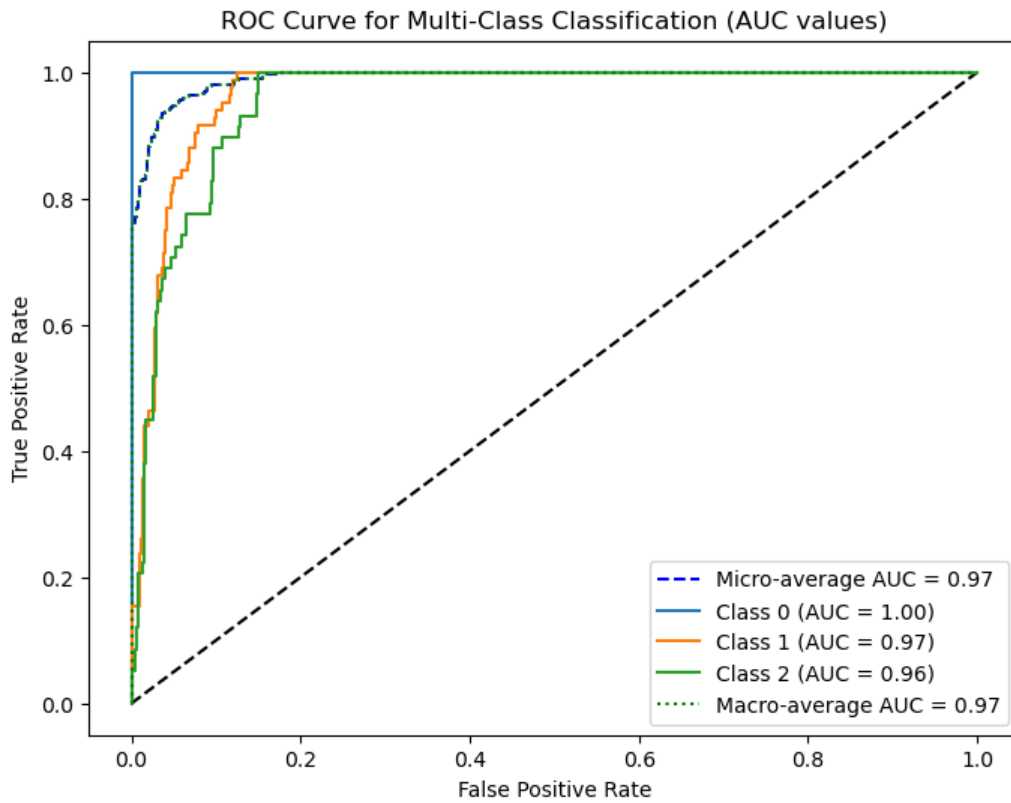


Figure 5.2: Receiver Operating Characteristic (ROC) curve for the first approach

Receiver Operating Characteristic (ROC) curve for the XGBoost model on an imbalanced dataset, depicting the trade-off between the true positive rate (sensitivity) and the false positive rate. Both the micro-averaged and macro-averaged AUC values are 0.97, indicating strong overall model, with higher values indicating better performance.

The following table summarizes the key performance metrics for the XGBoost classifier, calculated using weighted averages to account for class imbalances.

Metric	Value
Accuracy	0.9287
Precision	0.9281
Recall	0.9287
F1 Score	0.9283

These metrics indicate that the model achieves a high degree of accuracy, precision, and recall across all classes. The weighted F1 score of 0.9283 further highlights

the balance between precision and recall, demonstrating that the model effectively minimizes both false positives and false negatives.

Results of the second approach

The confusion matrix (Figure 5.3) shows the model's performance across the three classes. The XGBoost model achieved a balanced performance, correctly classifying all eight samples from Class 0 (control group). For Class 1 (SCZ) and Class 2 (BD), two and one samples, respectively, were misclassified, indicating slight challenges in distinguishing between these classes.

Overall, the model achieved:

Metric	Value
Accuracy	0.90
Precision	0.90
Recall	0.90
F1 Score	0.90

These metrics suggest a strong overall performance, though minor imbalances in classification effectiveness remain, particularly between SCZ and BD groups.

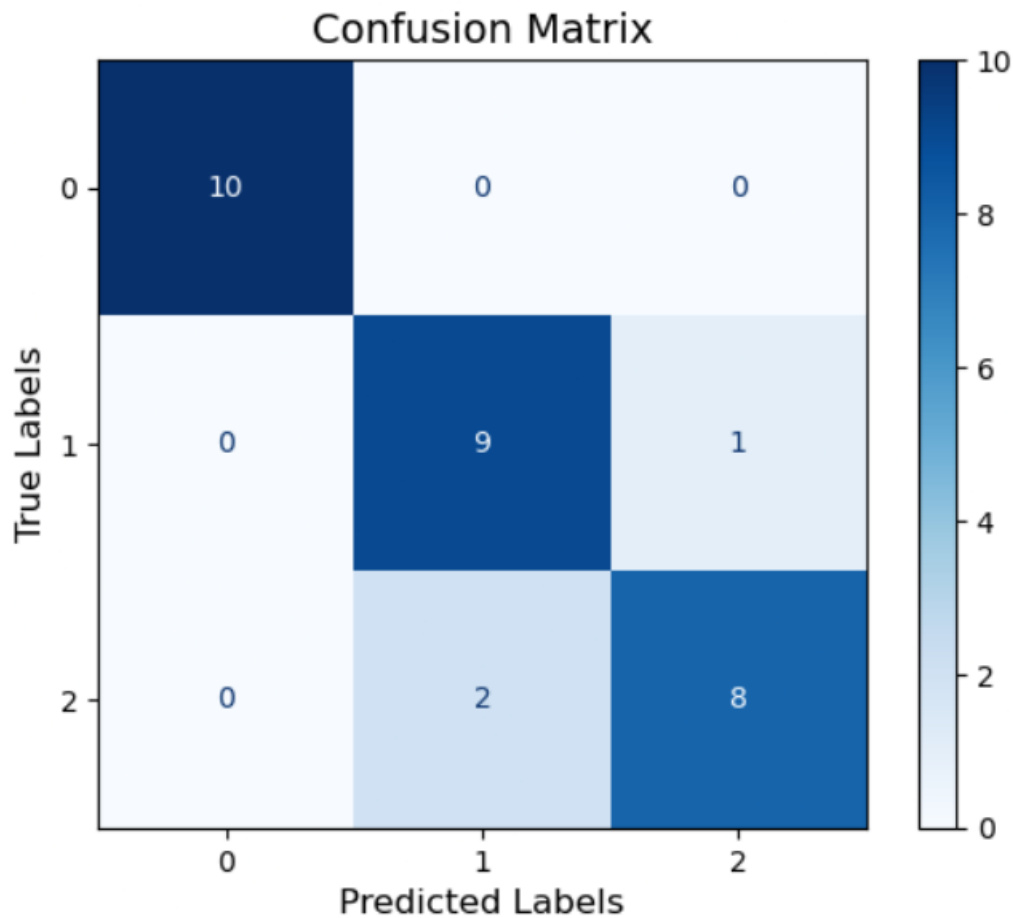


Figure 5.3: Confusion Matrix for the second approach

Confusion matrix for the XGBoost model on the balanced dataset, illustrating classification performance. The x- and y-axes represent the control (0), schizophrenia (1), and bipolar disorder (2) groups. Diagonal elements, from the top left to the bottom right, indicate correctly classified instances, while off-diagonal elements represent misclassifications.

The ROC curves (Figure 5.4) highlight the separability of classes. The AUC scores for Class 0, Class 1, and Class 2 were 1.00, 0.94, and 0.94, respectively, with macro- and micro-averages of 0.97 and 0.97. The high AUC values indicate strong discrimination capability for the control group and good separability for the SCZ and BD classes. However, slight overlaps in the feature space may have contributed to the observed misclassification.

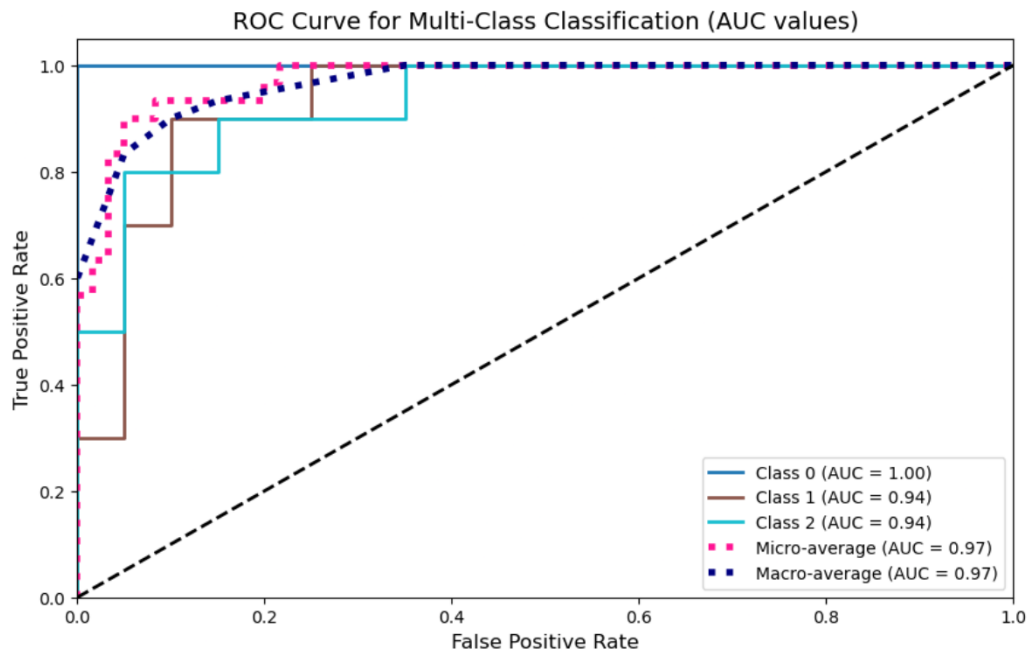


Figure 5.4: Receiver Operating Characteristic (ROC) curve for the second approach

Receiver Operating Characteristic (ROC) curve for the XGBoost model on the balanced dataset, depicting the trade-off between the true positive rate (sensitivity) and the false positive rate. Both the micro-averaged and macro-averaged AUC values are 0.97, indicating strong overall model, with higher values indicating better performance.

This approach demonstrates significant performance improvements, achieving high accuracy and strong AUC values. However, reducing the dataset size comes at the cost of generalisability.

5.2 Insights gained from the results

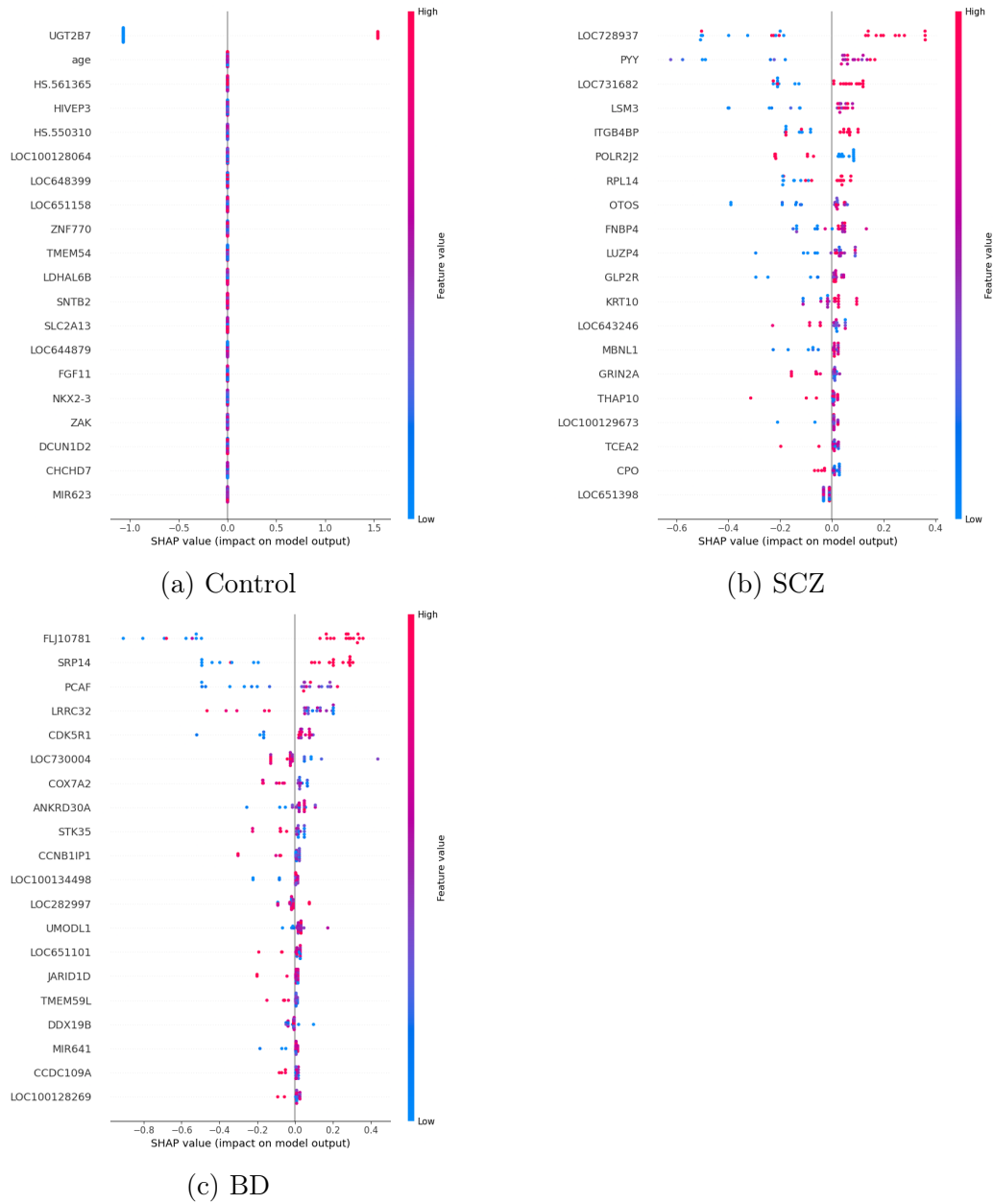


Figure 5.5: Feature Importance

SHAP-based feature importance visualization for the XGBoost model for multi-class classification. The plots illustrate the impact of individual features (genes) on the classification processes. Feature values are color-coded, with red indicating high values and blue indicating low values.

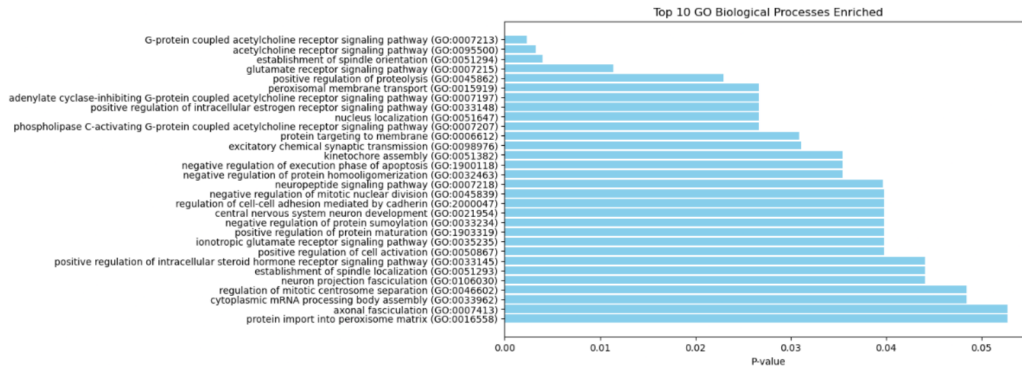


Figure 5.6: Biological Pathway Analysis

Top 10 enriched Gene Ontology (GO) biological processes based on functional enrichment analysis. The x-axis represents the p-value, indicating statistical significance, while the y-axis lists the enriched biological processes. Lower p-values suggest stronger enrichment.

Functional enrichment and network analyses were performed to investigate the biological relevance of the 90 most important genes identified through SHAP feature importance analysis.

Gene Ontology (GO) enrichment analysis

To identify the overrepresented biological processes, **Gene Ontology (GO) enrichment analysis** [12] was performed using the *Enrichr* [9] tool. The gene list was submitted to the *GO_Biological_Process_2018* database, and the results were retrieved in a tab-separated format. The top 30 enriched GO terms were filtered according to statistical significance ($p\text{-value} < 0.5$), and a bar chart was generated to visualize the most relevant biological pathways (shown in Figure 5.6).

Interpretation of GO Findings

GO enrichment analysis revealed several biological processes significantly associated with the selected genes and corresponding p-values. Highly enriched processes were selected based on criteria emphasizing that lower p-values signify greater statistical significance. As illustrated in Figure 5.6, GO enrichment analysis revealed significant enrichment in pathways such as **G-protein coupled acetylcholine receptor signalling pathway** (GO:0007213, $p = 0.0023$), **central nervous system neuron development** (GO:0021954, $p = 0.0397$), **establishment of spindle orientation** (GO:0051294, $p = 0.0039$) and **glutamate receptor signalling pathway** (GO:0007215, $p = 0.011$).

G-protein coupled acetylcholine receptors, also known as **muscarinic**

acetylcholine receptors (mAChRs), are a type of *G-protein coupled receptor (GPCR)*. According to the literature [[49],[6]], it is a membrane-embedded protein expressed in the central nervous system (CNS) and other tissues in the periphery body. It plays a crucial role in regulating various physiological processes. Muscarinic receptors play an important role in neuronal functions, including regulating the dopaminergic system, which is responsible for various cognitive and motor functions. Imbalances in this system have been implicated in cognitive dysfunction and mood regulation, which is common in both schizophrenia and bipolar disorder, supporting the role of these processes in psychiatric disorders [6]. The genes involved are CHRM4 and CDK5R1. According to information gathered from the NCBI Gene database, CHRM4 (muscarinic acetylcholine receptor M4) is specifically linked to attention and memory deficits in schizophrenia [41]. CDK5R1 has neurodevelopmental functions, and its dysregulation could affect synaptic plasticity [40].

Central Nervous System Neuron Development is a biological process involving the **NPY (neuropeptide Y)** gene. The NPY is integral in the development of neurons in the central nervous system (**neurogenesis and neuronal migration**) and is involved in psychiatric disorders such as SCZ and BD [47], [39]. Altered neurodevelopment is a well-known feature of SCZ, contributing to brain structure abnormalities, synaptic dysfunction, and impaired connectivity. BD also shows signs of disrupted neurodevelopment, but the timing and severity of these changes may vary throughout an individual's life compared to SCZ. The findings reveal that NPY mRNA levels are significantly reduced in the frontal cortex of individuals with SCZ and BD, with this reduction is more pronounced in the frontal cortex compared to the temporal cortex [30]. Furthermore, genetic studies indicate that genes related to NPY are more impaired in affective disorders like SCZ and BD. However, NPY dysfunction remains a significant feature of schizophrenia, particularly affecting GABAergic interneurons, further highlighting its role in the disorder's neurobiological underpinnings [54], [7], [30].

The enrichment of genes **KIF25** and **CENP-A**, which are involved in **spindle orientation during cell division**. The orientation of the spindle fibers during cell division helps determine how many neurons will be produced and how the brain will grow, suggesting a potential role in neurodevelopmental processes [31]. Proper **mitotic spindle orientation** is crucial for maintaining **chromosomal stability** and **ensuring accurate cell division**, particularly in neural progenitor cells. Disruptions in these pathways may contribute to **abnormal brain development**, which has been implicated in psychiatric disorders such as schizophrenia [35].

The glutamate receptor signalling pathway involves the genes **GRIN2A** and **CDK5R1**. **GRIN2A** encodes a subunit of the *N-methyl-D-aspartate (NMDA) receptor*, a key component in **glutamatergic synaptic transmission and plasticity**. Hypofunction of the NMDA receptor has been implicated in many psychiatric disorders, including schizophrenia [17], contributing to cognitive impairments and altered neural connectivity [42], [26]. Additionally, **CDK5R1** [40], which regulates

neuronal signalling and synaptic development, plays a pivotal role in maintaining proper glutamatergic function. Dysregulation of this pathway highlights the overlap between the neurochemical mechanisms underlying schizophrenia and bipolar disorder, while its prominent association with NMDA receptor dysfunction underscores its greater relevance for schizophrenia [40], [4]. These findings suggest that therapeutic strategies targeting glutamatergic signalling may hold promise for addressing the core symptoms, particularly schizophrenia.

Protein-Protein Interaction (PPI) network

To explore interactions among identified genes, a **Protein-Protein Interaction (PPI) network** was constructed using data retrieved from the *STRING database* [50]. The gene list was formatted and submitted to the *STRING API endpoint*, specifying *Homo sapiens* (species ID: 9606) as the target organism. The interaction data was collected in a tabular format, including protein identifiers (stringId_A and stringId_B), recognizable protein names (preferredName_A and preferredName_B) interaction scores, and the overall confidence score for each interaction.

The network graph was then generated using *NetworkX* [19], where the nodes represent the proteins and the edges represent functional interactions.

To highlight **highly connected hub proteins**, **degree centrality analysis** was performed, and proteins with a centrality score greater than 0.2 were visually distinguished within the network. The graph is visualized in Figure 5.7, showcasing the identified proteins' interaction patterns.

These analyses provide insights into the biological functions and molecular interactions of the most important genes in distinguishing SCZ, BD, and healthy individuals as identified by the machine learning model.

Protein-Protein Interaction Analysis on Neuropsychiatric Disorders

To further investigate the molecular mechanisms underlying schizophrenia and bipolar disorder, a protein-protein interaction (PPI) analysis was conducted using the *STRING database*. The data retrieved from the *STRING database* consisted of high-confidence interactions between proteins identified as significant in the Gene Ontology (GO) enrichment analysis. These interactions provide insights into the functional relationships among key proteins and their potential roles in neurodevelopmental and neurotransmitter signalling pathways.

Among the identified interactions, several involved proteins associated with **synaptic transmission, neuronal development, and neurotransmitter regulation**. These processes are known to be dysregulated in schizophrenia and bipolar disorder [18]. Notably, **NPY (Neuropeptide Y)**, a gene implicated in **central nervous system (CNS) neuron development (GO:0021954)**, was found to interact with **PYY (Peptide YY)**. This interaction had a high confidence score

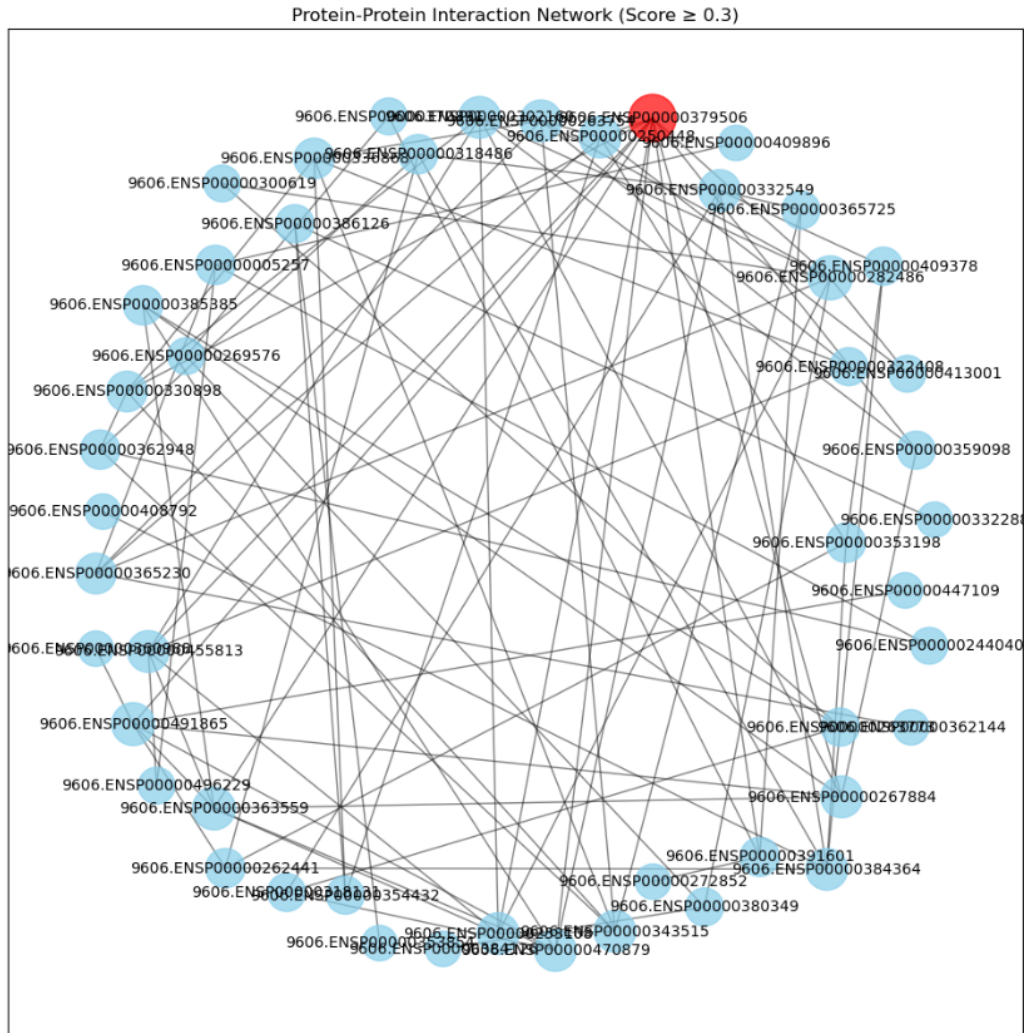


Figure 5.7: Protein-Protein Interaction (PPI) network highlighting interactions between identified proteins with a minimum score threshold of 0.3

Protein-Protein Interaction (PPI) network graph illustrating interactions between identified proteins, with a minimum interaction score threshold of ≥ 0.3 . Nodes represent proteins, while edges indicate predicted interactions. Node colors reflect different significance levels or interaction strengths, with highlighted nodes potentially representing key hub proteins.

(0.98), suggesting a strong functional association. Neuropeptides such as **NPY** possess critical roles in mood regulation, stress response, and synaptic plasticity by modulating anxiety and depression-related behaviours, influencing synaptic function, and interacting with stress and immune systems [56, 20]. Their dysregulation has been observed in both schizophrenia and bipolar disorder [38]. This finding supports the hypothesis that neuropeptide signalling alterations contribute to these disorders' pathophysiology.

In addition, interactions related to **glutamatergic signalling pathways** were identified. **GRIN2A**, previously linked to the **glutamate receptor signaling pathway (GO:0007215, GO:0035235)**, was not directly present in the retrieved interactions. However, related proteins, including **EIF3G**, **EIF6**, and **RPL14**, exhibited strong interactions, with **EIF6** and **RPL14** demonstrating a particularly high confidence score (0.996). The presence of these interactions suggests that translation regulation mechanisms, possibly affecting glutamate receptor subunits [43], [27], [28], could play a role in modulating the synaptic transmission deficits observed in schizophrenia. Given that **NMDA receptor hypofunction** is a well-documented feature of schizophrenia [44], the involvement of translation-associated proteins in the PPI network highlights a potential regulatory mechanism influencing NMDA receptor expression or function.

Further, the interactions involving **FOXA1 (Forkhead Box A1)** and **KAT2B (Lysine Acetyltransferase 2B)** suggest a link between **epigenetic regulation and neurodevelopment**. **FOXA1** is associated with **intracellular estrogen receptor signalling (GO:0033148)**, a pathway that has been implicated in gene-specific vulnerabilities to psychiatric disorders [14], [23]. The interaction between **FOXA1** and **KAT2B (score: 0.565)** may indicate a transcriptional regulatory mechanism influencing gene expression in neuronal differentiation and synaptic plasticity, disrupted processes in schizophrenia and bipolar disorder.

Notably, **CENP-A (Centromere Protein A)**, previously identified in GO enrichment under **establishment of spindle orientation (GO:0051294)**, was found to interact with **KAT2B (score: 0.633)**. Since abnormalities in mitotic processes and neuronal proliferation have been suggested as potential developmental risk factors for schizophrenia, [52] this interaction provides further support for the neurodevelopmental hypothesis of schizophrenia, where dysregulated cell cycle processes contribute to abnormal brain connectivity.

The peroxisomal transport protein **PEX5**, previously linked to **peroxisomal membrane transport (GO:0046913)**, was not found in the retrieved PPI data. However, related processes involved in protein transport, such as interactions between **SRP14** and **COX7A2 (score: 0.639)**, suggest possible roles for peroxisomal or mitochondrial dysfunction in neuropsychiatric disorders. Emerging evidence indicates that oxidative stress and mitochondrial dysfunction contribute to schizophrenia and bipolar disorder, and the presence of these interactions aligns with this hypothesis [29].

The integration of **PPI data with GO enrichment analysis** highlights sev-

eral biologically relevant pathways that may underlie the molecular mechanisms of schizophrenia and bipolar disorder. Key findings include:

1. **Neuropeptide interactions (NPY-PYY)** reinforcing the role of **neuropeptide signaling** in mood and cognitive regulation.
2. **EIF6-RPL14 interactions** suggesting potential regulatory mechanisms influencing **glutamatergic neurotransmission**.
3. **FOXA1-KAT2B and CENPA interactions** linking **epigenetic and neurodevelopmental pathways** to psychiatric disorders.
4. **Mitochondrial and protein transport interactions**, indicating possible **metabolic dysfunction** in schizophrenia and bipolar disorder.

These findings provide new avenues for further research into the molecular aetiology of these disorders and may inform targeted therapeutic interventions focusing on neurotransmitter modulation, neurodevelopmental regulation, and metabolic stability.

6. Conclusion

6.1 Summary of Findings

This study demonstrated the utility of ML models, specifically XGBoost, in classifying SCZ and BD using gene expression data. The model achieved high accuracy (92.87% on the imbalanced dataset, 90% on the balanced dataset), effectively distinguishing between SCZ, BD, and healthy controls. SHAP-based feature importance analysis identified 90 genes, with functional enrichment analysis linking them to key neurobiological pathways, such as glutamate receptor signaling and neuropeptide activity. Protein-protein interaction analysis further elucidated molecular mechanisms potentially underlying these psychiatric disorders.

6.2 Contributions

This research makes several key contributions to psychiatric genetics and ML-based diagnostics:

- Developed a robust multiclass classification model to distinguish between SCZ and BD using genetic data.
- Demonstrated the effectiveness of boosting algorithms in handling imbalanced biomedical datasets.
- Provided insights into biologically relevant genes and pathways, potentially informing future biomarker discovery.
- Integrated ML-based feature selection with functional enrichment and network analysis to enhance the interpretability of results.

6.3 Implications

The study underscores the potential of ML techniques in psychiatric diagnosis, paving the way for more objective, data-driven assessments. By identifying key genetic markers associated with SCZ and BD, this approach may contribute to early diagnosis and personalized treatment strategies. Additionally, the integration of feature selection and pathway enrichment analysis could aid in the development of targeted therapeutic interventions.

6.4 Limitations

Despite its promising results, this study has several limitations:

- The dataset size remains a constraint, particularly for minority classes, which may impact model generalizability.
- Gene expression data derived from peripheral blood may not fully capture neurobiological processes occurring in the brain.
- Potential Ethnic and Population biases may arise due to the absence of diverse population cohorts in the training data.
- While XGBoost performed well, additional ML techniques, including deep learning, could be explored for enhanced performance.

6.5 Future Work

To build upon these findings, future research should:

- Expand the dataset to include larger and more diverse cohorts to improve model generalizability.
- Investigate additional ML approaches, such as deep learning and ensemble methods, to enhance classification accuracy.
- Validate findings using independent datasets and functional studies to confirm biological relevance.
- Explore integration with multimodal data, including neuroimaging and clinical assessments, for a more comprehensive diagnostic model.

In conclusion, this study demonstrates the potential of ML-driven genomic analysis for improving psychiatric diagnostics, providing a foundation for future research in precision medicine and biomarker discovery for SCZ and BD.

Author's Statement: I hereby expressly declare that, according to the article 8 of Law 1559/1986, this dissertation is solely the product of my personal work, does not infringe any intellectual property, personality and personal data rights of third parties, does not contain works/contributions from third parties for which the permission of the authors/beneficiaries is required, is not the product of partial or total plagiarism, and that the sources used are limited to the literature references alone and meet the rules of scientific citations.

Bibliography

- [1] Saghir Ahmed, Basit Raza, Lal Hussain, Amjad Aldweesh, Abdulfattah Omar, Mohammad Shahbaz Khan, Elsayed Tageldin, and Muhammad Nadim. The deep learning resnet101 and ensemble xgboost algorithm with hyperparameters optimization accurately predict the lung cancer. *Applied Artificial Intelligence*, 37, 06 2023.
- [2] Rosa Lundbye Allesøe, Wesley K. Thompson, Jonas Bybjerg-Grauholm, David M. Hougaard, Merete Nordentoft, Thomas Werge, Simon Rasmussen, and Michael Eriksen Benros. Deep learning for cross-diagnostic prediction of mental disorder diagnosis and prognosis using danish nationwide register and genetic data. *JAMA Psychiatry*, 80(2):146–155, December 2022. Accepted for Publication: October 12, 2022. Published Online: December 7, 2022.
- [3] Ana C. Andreazza, Li Shao, Jun-Feng Wang, and L. Trevor Young. Mitochondrial complex i activity and oxidative damage to mitochondrial proteins in the prefrontal cortex of patients with bipolar disorder. *Archives of General Psychiatry*, 67(4):360–368, 04 2010.
- [4] D. T. Balu. The nmda receptor and schizophrenia: From pathophysiology to treatment. *Advances in Pharmacology*, 76:351–382, 2016. Epub 2016 Mar 4. PMID: 27288082.
- [5] Mahesh Batta. Machine learning algorithms - a review. *International Journal of Science and Research (IJSR)*, 9(1), 2020. ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426.
- [6] Andreas Bock, Ramona Schrage, and Klaus Mohr. Allosteric modulators targeting cns muscarinic receptors. *Neuropharmacology*, 136:427–437, 2018. Neuropharmacology on Muscarinic Receptors.
- [7] Robert W Buchanan, Katalin Vadar, Patrick E Barta, and Godfrey D Pearson. Structural evaluation of the prefrontal cortex in schizophrenia. *American Journal of Psychiatry*, 155(8):1049–1055, 1998.
- [8] Alexandre Bureau. Rna expression in immortalized lymphocytes of schizophrenia and bipolar disorders patients and their unaffected relatives. <https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-8018>, 2019.
- [9] Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R Clark, and Avi Ma’ayan. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC bioinformatics*, 14:1–14, 2013.

- [10] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.
- [11] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.
- [12] The Gene Ontology Consortium. The gene ontology resource: enriching a gold mine. *Nucleic Acids Research*, 49(D1):D325–D334, 2021.
- [13] The International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460:748–752, 2009. Received 11 February 2009; Accepted 08 June 2009; Published 01 July 2009; Issue Date 06 August 2009.
- [14] Amanda Crider and Anilkumar Pillai. Estrogen signaling as a therapeutic target in neurodevelopmental disorders. *Journal of Pharmacology and Experimental Therapeutics*, 360(1):48–58, 2017.
- [15] Noémie Drancourt, Bruno Etain, Mohamed Lajnef, Chantal Henry, Aurélie Raust, Barbara Cochet, Flavie Mathieu, Sébastien Gard, Katia MBailara, L Zanouy, et al. Duration of untreated bipolar disorder: missed opportunities on the long road to optimal treatment. *Acta Psychiatrica Scandinavica*, 127(2):136–144, 2013.
- [16] Ian Ellison-Wright and Ed Bullmore. Anatomy of bipolar disorder and schizophrenia: A meta-analysis. *Schizophrenia Research*, 117(1):1–12, 2010.
- [17] Xuelai Fan, Wu Yang Jin, and Yu Tian Wang. The nmda receptor complex: a multifunctional machine at the glutamatergic synapse. *Frontiers in Cellular Neuroscience*, 8, 2014. Published: 10 June 2014.
- [18] Laura J Gray, Brian Dean, Helena C Kronsbein, Phillip J Robinson, and Elizabeth Scarr. Region and diagnosis-specific changes in synaptic proteins in schizophrenia and bipolar i disorder. *Psychiatry research*, 178(2):374–380, 2010.
- [19] Aric Hagberg, Pieter J Swart, and Daniel A Schult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Laboratory (LANL), Los Alamos, NM (United States), 2008.
- [20] Markus Heilig. The npy system in stress, anxiety and depression. *Neuropeptides*, 38(4):213–224, 2004.

- [21] Robert MA Hirschfeld, Lydia Lewis, Lana A Vornik, et al. Perceptions and impact of bipolar disorder: how far have we really come? results of the national depressive and manic-depressive association 2000 survey of individuals with bipolar disorder. *Journal of Clinical Psychiatry*, 64(2):161–174, 2003.
- [22] Oliver D. Howes and Shitij Kapur. The dopamine hypothesis of schizophrenia: Version iii—the final common pathway. *Schizophrenia Bulletin*, 35(3):549–562, 03 2009.
- [23] Wu Jeong Hwang, Tae Young Lee, Nahrie Suk Kim, and Jun Soo Kwon. The role of estrogen receptors and their signaling across psychiatric disorders. *International journal of molecular sciences*, 22(1):373, 2020.
- [24] IHME, Global Burden of Disease. Bipolar Disorder Prevalence: Estimated age-standardized prevalence per 100 people. Major processing by Our World in Data, 2024. Last updated: May 20, 2024. Next expected update: May 2028. Data range: 1990–2021. Unit:
- [25] IHME, Global Burden of Disease. Schizophrenia Prevalence: Estimated age-standardized prevalence per 100 people. Major processing by Our World in Data, 2024. Last updated: May 20, 2024. Next expected update: May 2028. Data range: 1990–2021. Unit:
- [26] Benjamin E. Jewett and Bicky Thapa. *Physiology, NMDA Receptor*. StatPearls Publishing, 2022. Last updated: December 11, 2022. Accessed: 29-Jan-2025.
- [27] Courtney F Jungers, Jonah M Elliff, Daniela S Masson-Meyers, Christopher J Phiel, and Sofia Origanti. Regulation of eukaryotic translation initiation factor 6 dynamics through multisite phosphorylation by gsk3. *Journal of Biological Chemistry*, 295(36):12796–12813, 2020.
- [28] Zaur M Kachaev, Sergey D Ivashchenko, Eugene N Kozlov, Lyubov A Lebedeva, and Yulii V Shidlovskii. Localization and functional roles of components of the translation apparatus in the eukaryotic cell nucleus. *Cells*, 10(11):3239, 2021.
- [29] Ines Khadimallah, Raoul Jenni, Jan-Harry Cabungcal, Martine Cleusix, Margot Fournier, Elidie Beard, Paul Klauser, Jean-François Knebel, Micah M Murray, Chrysa Retsa, et al. Mitochondrial, exosomal mir137-cox6a2 and gamma synchrony as biomarkers of parvalbumin interneurons, psychopathology, and neurocognition in schizophrenia. *Molecular Psychiatry*, 27(2):1192–1204, 2022.

- [30] J. Kuromitsu, A. Yokoi, T. Kawai, T. Nagasu, T. Aizawa, S. Haga, and K. Ikeda. Reduced neuropeptide y mrna levels in the frontal cortex of people with schizophrenia and bipolar disorder. *Brain research. Gene expression patterns*, 1 1:17–21, 2001.
- [31] Madeline A Lancaster and Juergen A Knoblich. Spindle orientation in mammalian cerebral cortical development. *Current opinion in neurobiology*, 22(5):737–746, 2012.
- [32] T. M. Laursen. Bipolar disorder, schizoaffective disorder, and schizophrenia overlap: A new comorbidity index. *Acta Psychiatrica Scandinavica*, 119(5):357–362, 2009. Online ahead of print: June 16, 2009.
- [33] Paul Lichtenstein, Benjamin H Yip, Camilla Björk, Yudi Pawitan, Tyrone D Cannon, Patrick F Sullivan, and Christina M Hultman. Common genetic determinants of schizophrenia and bipolar disorder in swedish families: A population-based study. *The Lancet*, 373(9659):234–239, 2009.
- [34] Ágnes Lublóy, Judit Lilla Keresztúri, Attila Németh, and Péter Mihalicza. Exploring factors of diagnostic delay for patients with bipolar disorder: a population-based cohort study. *BMC psychiatry*, 20:1–17, 2020.
- [35] DJ MacIntyre, DHR Blackwood, DJ Porteous, BS Pickard, and Walter J Muir. Chromosomal abnormalities and mental illness. *Molecular psychiatry*, 8(3):275–287, 2003.
- [36] H. K. Manji, J. A. Quiroz, J. L. Payne, J. Singh, B. P. Lopes, J. S. Viegas, and C. A. Zarate. The underlying neurobiology of bipolar disorder. *World Psychiatry: Official Journal of the World Psychiatric Association (WPA)*, 2(3):136–146, 2003.
- [37] Robert A. McCutcheon, Tiago Reis Marques, and Oliver D. Howes. Schizophrenia—an overview. *JAMA Psychiatry*, 77(2):201–210, 02 2020.
- [38] Julio César Morales-Medina, Yvan Dumont, and Rémi Quirion. A possible role of neuropeptide y in depression and stress. *Brain research*, 1314:194–205, 2010.
- [39] Shunsuke Morosawa, S. Iritani, H. Fujishiro, Hirotaka Sekiguchi, Youta Torii, Chikako Habuchi, K. Kuroda, K. Kaibuchi, and N. Ozaki. Neuropeptide y neuronal network dysfunction in the frontal lobe of a genetic mouse model of schizophrenia. *Neuropeptides*, 62:27–35, 2017.
- [40] National Center for Biotechnology Information. CDK5R1 cyclin dependent kinase 5 regulatory subunit 1 (Homo sapiens), 2025. Updated: 4-Jan-2025.

- [41] National Center for Biotechnology Information. CHRM4 cholinergic receptor muscarinic 4 (Homo sapiens), 2025. Accessed: 29-Jan-2025.
- [42] National Center for Biotechnology Information. GRIN2A glutamate ionotropic receptor NMDA type subunit 2A (Homo sapiens), 2025. Accessed: 29-Jan-2025.
- [43] Kisho Obi-Nagata, Yusuke Temma, and Akiko Hayashi-Takagi. Synaptic functions and their disruption in schizophrenia: From clinical evidence to synaptic optogenetics in an animal model. *Proceedings of the Japan Academy, Series B*, 95(5):179–197, 2019.
- [44] John W Olney, John W Newcomer, and Nuri B Farber. Nmda receptor hypo-function model of schizophrenia. *Journal of psychiatric research*, 33(6):523–533, 1999.
- [45] World Health Organization. Bipolar disorder. July 2024. Accessed: 2024-01-09.
- [46] Deborah A Perlick, Robert A Rosenheck, John F Clarkin, Paul K Maciejewski, JoAnne Sirey, Elmer Struening, and Bruce G Link. Impact of family burden and affective response on clinical outcome among patients with bipolar disorder. *Psychiatric services*, 55(9):1029–1035, 2004.
- [47] J. Redrobe, Y. Dumont, and R. Quirion. Neuropeptide y (npy) and depression: from animal studies to the human condition. *Life sciences*, 71 25:2921–37, 2002.
- [48] Teshome Shibre, Derege Kebede, Atalay Alem, Aleyamehu Negash, N Deyassa, A Fekadu, Daniel Fekadu, Lars Jacobsson, and Gunnar Kullgren. Schizophrenia: illness impact on family members in a traditional society–rural ethiopia. *Social Psychiatry and Psychiatric Epidemiology*, 38:27–34, 2003.
- [49] Jeffrey S. Smith, Ari S. Hilibrand, Meredith A. Skiba, Andrew N. Dates, Victor G. Calvillo-Miranda, and Andrew C. Kruse. The m3 muscarinic acetylcholine receptor can signal through multiple g protein families. *Molecular Pharmacology*, 105(6):386–394, June 2024.
- [50] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, et al. The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research*, 51(D1):D638–D646, 2023.
- [51] Yannis J. Trakadis, Sameer Sardaar, Anthony Chen, Vanessa Fulginiti, and Ankur Krishnan. Machine learning in schizophrenia genomics, a case-control

- study using 5,090 exomes. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 2018.
- [52] Nadia Tsankova, William Renthall, Arvind Kumar, and Eric J Nestler. Epigenetic regulation in psychiatric disorders. *Nature Reviews Neuroscience*, 8(5):355–367, 2007.
- [53] K. Wahlbeck, J. Westman, M. Nordentoft, M. Gissler, and T. M. Laursen. Outcomes of nordic mental health systems: life expectancy of patients with mental disorders. *The British journal of psychiatry: the journal of mental science*, 199(6):453–458, 2011.
- [54] CS Weickert, TM Hyde, BK Lipska, MM Herman, DR Weinberger, and JE Kleinman. Reduced brain-derived neurotrophic factor in prefrontal cortex of patients with schizophrenia. *Molecular psychiatry*, 8(6):592–610, 2003.
- [55] World Health Organization. Schizophrenia, 2022.
- [56] Gang Wu, Adriana Feder, Gregers Wegener, Christopher Bailey, Shireen Saxena, Dennis Charney, and Aleksander A Mathé. Central functions of neuropeptide y in mood and anxiety disorders. *Expert opinion on therapeutic targets*, 15(11):1317–1331, 2011.
- [57] Qingxia Yang, Qiaowen Xing, Qingfang Yang, and Yaguo Gong. Classification for psychiatric disorders including schizophrenia, bipolar disorder, and major depressive disorder using machine learning. *Computational and Structural Biotechnology Journal*, 20:5054–5064, 2022.