ROLE OF OMIC DATA IN THE IDENTIFICATION OF SCHIZOPHRENIA DISORDER PATIENTS

ANNOTATED BIBLIOGRAPHY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS OF THE DEGREE OF

BACHELOR OF THE SCIENCE IN ENGINEERING

**Submitted by:**

Aarthy V. (2019/E/001)

Senthooran S. (2019/E/132)

**DEPARTMENT OF COMPUTER ENGINEERING**

**FACULTY OF ENGINEERING**

**UNIVERSITY OF JAFFNA**

AUGUST 2023

ROLE OF OMIC DATA IN THE IDENTIFICATION OF SCHIZOPHRENIA DISORDER PATIENTS

Supervisor: Dr. P. Jeyananthan

Examination Committee:

Lecturer 1: ……………………….....

Lecturer 2: …………………………..

CONTRIBUTION TO THE PROPOSAL BY THE MEMBERS IN THE GROUP

|  |  |  |
| --- | --- | --- |
| Sections | 2019/E/001 | 2019/E/132 |
| CHAPTER 1: INTRODUCTION |  |  |
| 1.1 Motivation and overview |  |  |
| 1.2 Aims and objectives |  |  |
| 1.3 Research scope |  |  |
| CHAPTER 2: LITERATURE REVIEW |  |  |
| 2.1 Introduction |  |  |
| 2.2 Research gap |  |  |
| 2.3 Forecasting model |  |  |
| 2.4 Performance analysis |  |  |
| 2.5 Available databases |  |  |
| CHAPTER 3: METHODOLOGY AND RESEARCH PLAN |  |  |
| 3.1 Methodology in brief |  |  |
| 3.2 Detailed methodology |  |  |
| 3.2.1 Breast cancer data selection |  |  |
| 3.2.2 Data preprocessing |  |  |
| 3.2.3 Feature selection |  |  |
| 3.2.4 Apply machine learning methods |  |  |
| 3.2.5 Compare performance |  |  |
| 3.3 Timeline |  |  |
| CHAPTER 4: PROGRESS TO DATE |  |  |
| 4.1 Literature review |  |  |
| 4.2 Database collection |  |  |
| 4.2.1 Phenotype data selection |  |  |
| 4.2.2 Dataset selection |  |  |
| 4.3 Database preparation |  |  |
| REFERENCE |  |  |
| APPENDIX |  |  |

Contents

[Content 4](#_Toc135863503)

[Abbreviations 5](#_Toc135863504)

[INTRODUCTION 6](#_Toc135863505)

[Motivation and overview 6](#_Toc135863506)

[Aims and objectives 6](#_Toc135863507)

[Research scope 7](#_Toc135863508)

[LITERATURE REVIEW 7](#_Toc135863509)

[Introduction 7](#_Toc135863510)

[Research gap 7](#_Toc135863511)

[Forecasting model 8](#_Toc135863512)

[Performance analysis 10](#_Toc135863513)

[Available databases 11](#_Toc135863514)

[METHODOLOGY AND RESEARCH PLAN 11](#_Toc135863515)

[Methodology in brief 11](#_Toc135863516)

[Detailed methodology 12](#_Toc135863517)

[Schizophrenia data selection 12](#_Toc135863518)

[Data preprocessing 13](#_Toc135863519)

[Feature selection 13](#_Toc135863520)

[Apply machine learning methods 14](#_Toc135863521)

[Compare performance 14](#_Toc135863522)

[Timeline 15](#_Toc135863523)

[PROGRESS TO DATE 15](#_Toc135863524)

[Literature review 15](#_Toc135863525)

[Database collection 15](#_Toc135863526)

[Phenotype data selection 15](#_Toc135863527)

[Data set selection 15](#_Toc135863528)

[Database preparation 15](#_Toc135863529)

[REFERENCES 16](#_Toc135863530)

# Abbreviations

ANN – Artificial neural network

AUC – Area under the curve

BP – bipolar disorder

CMC - Common Mind Consortium

CoRSIV - correlated regions of systematic interindividual epigenetic variation

DLPFC - Dorsolateral prefrontal cortex

DNA – Deoxyribonucleic acid

GB – Gradient boosting

GWAS – Genome-wide association studies

KNN – K-nearest neighbor

MDD – Major depressive disorder

ML – Machine learning

mRNA – Micro RNA

PGC - Psychiatric Genomics Consortium

RF – Random Forest

RNA – Ribonucleic acid

SCZ – Schizophrenia

SDST - Schizophrenia Data and Software Tool

SVM – Support vector machine

# INTRODUCTION

## Motivation and overview

SCZ is a severe mental disorder that affects a significant portion of the global population, with a prevalence rate of approximately 1%, and generally appears in subjects aged 15 to 25 years [5,8]. Nearly 5% of persons with schizophrenia die from suicide, and 25–50% of these patients attempt suicide at some point in their lifespan [17]. The symptoms of SCZ are currently classified into three categories: positive symptoms, negative symptoms, and cognitive impairments [12]. Negative symptoms like social isolation and apathy are seen to be significant contributors to the patient's performance and quality of life, despite the fact that positive symptoms like delusions and hallucinations are the main characteristics of schizophrenia [13]. Unfortunately, the current diagnostic procedures for schizophrenia are limited, often resulting in delayed and inaccurate diagnoses [10].

SCZ is a complex psychiatric disorder with a multifactorial etiology involving genetic, environmental, and neurobiological factors [14]. Accurate prediction of schizophrenia onset is challenging due to its heterogeneous nature and the lack of specific diagnostic biomarkers. However, omic data, including genomics (genetic variants, copy number variations), transcriptomics (gene expression), epigenomics (DNA methylation, chromatin, and histone modification), proteomics (protein expression), and metabolomics (metabolite expression), have emerged as valuable resources for unraveling the molecular mechanisms underlying schizophrenia.

The majority of earlier studies concentrated on categorizing psychiatric diseases including SCZ, BP, and MDD using various machine learning algorithms such as SVM, RF, and K-NN, and comparing their accuracy scores [8,10]. In addition, for the prediction of SCZ, several studies concentrated on just one form of omic data [3,5,9]. Genome-wide association studies (GWAS), differential gene expression, and differential methylation analyses are examples of conventional single omics study methods [15]. Others concentrated on non-omic data as well, such as demographic characteristics and event-related potentials [1,4].

Machine learning techniques have the potential to revolutionize the diagnosis and management of schizophrenia. By leveraging the power of computational algorithms and analyzing diverse data sources, ML models can improve early detection, diagnostic accuracy, and personalized treatment approaches. Implementing ML-based diagnostic tools can lead to earlier interventions, better patient outcomes, and reduced burden on individuals affected by these debilitating mental disorders. At the end of this study, we intend to examine which omic data significantly influence SCZ prediction for various ML algorithms.

## Aims and objectives

The studies conclusively show that machine learning techniques are more efficient compared to psychiatrists' diagnoses following comprehensive clinical examinations. The aim of our study is to investigate the role of omic data in the identification and prediction of schizophrenia disorder patients. By selecting and analyzing specific genomic markers and omic data types to determine how it can be utilized to improve the accuracy of schizophrenia prediction.

Our objective is to develop a comprehensive machine learning methodology for analyzing omic data, for early detection of schizophrenia. The study provides a combination of machine learning techniques for pre-processing, feature selection, and prediction. Also, we aim to study the factors that could be influence on the performance of machine learning methods on each omic data. Furthermore, this study highlights the potential impact of incorporating omic data in the field of schizophrenia research which supports clinicians in making informed decisions and predictions.

## Research scope

The scope of our work is to determine how omic data may be used to predict schizophrenia. This methodology will make it easier to determine which genomic data is most helpful for predicting schizophrenia and offering therapy before the condition worsens. By doing this, millions of patients around the world would benefit. Additionally, the medical industry will find great value for this study.

# LITERATURE REVIEW

## Introduction

Schizophrenia is a complex mental disorder characterized by a range of emotional abnormalities such as hallucinations, delusions, disorganized thinking, and social withdrawal. Identifying individuals with schizophrenia at an early stage is important for providing good treatment and support. In recent years, the field of psychiatric research has witnessed a significant expansion in the application of omic data, including genomics, transcriptomics, proteomics, and metabolomics, to unravel the underlying mechanisms and identify potential biomarkers associated with schizophrenia.

This literature review aims to explore the role of omic data in the identification of schizophrenia patients and, it will discuss the existing studies that have utilized non-omic data sources, such as clinical assessments, neuroimaging, event-related potentials, and cognitive evaluations. By reviewing these studies, we can gain insights into the advancements made and the potential limitations encountered in the use of omic and non-omic data for identifying schizophrenia patients. This understanding will lay the foundation for exploring the potential benefits and challenges of incorporating omic data into the identification and characterization of schizophrenia, ultimately contributing to the development of more precise diagnostic approaches and personalized treatment strategies.

## Research gap

The accuracy of various ML algorithms in the prediction of hospitalized patients with SCZ was compared by authors in [6]. The accuracy of the decision tree, the adaboost, the random forest, the naive bayes, the support vector machine, and the k nearest neighbor were all compared. And they achieved the best accuracy of 72.7% for the RF algorithm with 79.6%, 72.8%, 72.7%, and 72.7% for AUC, precision, F1 score, and recall respectively. Even though, 72.7% accuracy is not enough for accurate prediction. Other than RF, Adaboost and decision tree showed 70.8% and 68.2% accuracy, respectively. Naïve byes report 67.0% accuracy, k-NN with 67.7% accuracy, and SVM with the lowest value of accuracy at 65.7%.

The author collected patient admission records from 11 public hospitals in Castilla and Leon, Spain between the years 2005 and 2015. A total of 6933 patients with various mental diseases are included in the data collection, including 3002 patients with schizophrenia and 3931 patients with other mental disorders. The 6933 patients have a total of 11,884 admission records, of which 5968 documents are for patients with schizophrenia and 5916 records are for patients with other mental illnesses. Males comprise 55.9% of the data set's records (6639 out of 11,884 total), while females comprise 44.14% of the records (5245 out of 11,884 total). And patients range in age from 13 to 97 years old. But the limitation of this work is that it included only the data of patients who are hospitalized, and people with SCZ from that region who are not hospitalized have not been included. And their best accuracy value is not enough.

So, in our study, we are going to compare the Gene expression RNA seq, DNA methylation, and microRNA seq data with different ML algorithms to predict the role of omic data.

## Forecasting model

Omic data is the term used to describe massive amounts of data produced by various high-throughput biological technologies that record detailed information about many features of biological systems. The term "omic" comes from the suffix "-omics" and refers to a variety of biological specialties, including genomics, transcriptomics, proteomics, metabolomics, and others. Each of these fields specializes in the investigation of a particular level of biological knowledge.

The study of an organism's entire DNA sequence, including all its genes and non-coding sections, is known as genomics. Information on the DNA sequence, genetic variants, and structural differences within or between populations is available through genomic data. The study of all RNA molecules (transcripts) present in a cell or tissue at a certain time is known as transcriptomics. It offers details on the patterns of gene expression, such as which genes are activated or inactive and how much RNA is generated by each gene. The study of all proteins expressed by a cell, tissue, or organism is at the heart of proteomics. Identification, quantification, and studies of protein changes, interactions, and functional functions are its main objectives. The entire collection of tiny molecules (metabolites) found in a biological sample is analyzed using metabolomics. It offers details on biochemical processes, metabolic pathways, and metabolite levels, which might reveal cellular functions and physiological states. The study of epigenomics focuses on the DNA and protein changes that control gene expression without altering the underlying DNA sequence. It covers the investigation of non-coding RNA molecules, chromatin structure, histone modifications, and DNA methylation.

In most of the past researchers, authors used different levels of gene expression for the prediction of SCZ [3,5,7,8]. In most cases, researchers focused only on single omic data in the prediction of SCZ [3,5]. And some authors focused on finding the best ML algorithm for the prediction [10]. And, some others focused on non-omic data such as event-related potentials, clinical data, and demographic features of humans for the prediction of SCZ [1,4].

There are so many studies based on the prediction of SCZ using single omic data. Numerous genetic variations linked to schizophrenia have been discovered by genome-wide association studies (GWAS). In these investigations, the entire genes (genome) of huge populations is analyzed to find common genetic variants that are more common in people with schizophrenia than in healthy controls. These genetic variants are frequently found in or close to genes that affect the immunological response, neurotransmission, and brain development. The precise genetic variations discovered using GWAS often have negligible individual effects, and their overall contribution to the risk of schizophrenia is still unclear. Transcriptomic research has shed light on the different gene expression patterns in schizophrenia patients' brains compared to healthy controls. These investigations have emphasized the dysregulation of genes related to immunological response, synaptic function, and brain development, pointing to possible molecular pathways causing the illness.

Schizophrenia relates to changes in protein expression and signaling pathways, according to proteomic research. The modifications of synaptic proteins, neurotransmitter receptors, and molecules involved in signal transduction and communication between neurons have been clarified by these results. Schizophrenia is one of many psychiatric illnesses for which epigenetic processes can affect gene expression. Schizophrenia patients' DNA methylation patterns and histone modifications were altered, according to epigenomic studies, shedding light on possible epigenetic causes of the illness.

In [8], authors used transcriptomic data to identify the key genes related to SCZ, MDD, and BD. They analyzed data and construct the model using SVM to classify the psychiatric diseases. The main limitation is that they did not compare different machine-learning algorithms. However, they achieved an AUC value of 0.71.

In studies such as [3], authors identified the potential genetic factors associated with SCZ by analyzing genes and non-coding RNA, especially long non-coding RNA expression data. They used RNA-seq data collected from DLPFC on 254 subjects and applied different machine-learning algorithms. The work included a small sample size because of the difficulty in collecting brain tissues.

In the study [5], authors used correlated regions of systematic interindividual epigenetic variation (CoRSIVs) methylation in blood DNA for classifying SCZ patients from healthy controls. In this research, SCZ patients were identified using a single omic dataset; performance between alternative omic datasets was not evaluated. They might omit the most accurate genomic data.

The authors of [6] have tested algorithms such as decision tree, adaboost, random forest, naïve bayes, support vector machine, and k-nearest neighbor in the prediction of hospitalized patients with SCZ. The experiment showed the best accuracy of 72.7% using RF. 0.708%, 0.682%, 0.677%, 0.670%, and 0.641% accuracy values for adaboost, decision tree, k-NN, naïve bayes, and SVM respectively.

The authors of [18] used 72 individuals, who equate to 48 and 24 schizophrenia patients and healthy controls, respectively, to test algorithms like SVM, Naive Bayes, Random Forest, and gradient boosting to predict patients with schizophrenia and healthy controls. Using SVM and Random Forest, the studies revealed an accuracy of 58.2% and 68.6%, respectively.

In [7] authors used RNA-seq data from the dorsolateral prefrontal cortex and amygdala for the expansion of SCZ gene network knowledge. They used RF and factor analysis on transcriptomic data to select the most informative genes. But the limitation is that the sample size is relatively small.

Another author [10] filtered 15 papers from 1243 papers and compare three algorithms, SVM, RF, and GB in the prediction of SCZ and BP with higher accuracy. RF seemed to be more effective than GB and SVM. In [19], the authors found that a sample of 21 characteristics yielded accuracy values for Decision Tree, SVM, k-NN, and Random Forest of 72.3%, 78.7%, 76.5%, and 85.1%, respectively. The authors state that the data set consisted of around 500 patient medical records, and cross-validation was employed.

In [2], authors identified the molecular and cellular alteration that occurred in the brain of individuals with SCZ by analyzing the expression of genes involved in the glutamatergic neurotransmitter system. The authors used the data collected from post-mortem brain tissue. The main limitation of this article is that they used post-mortem brain tissue samples and the cause of death might not be SCZ.

Researchers have been involved in several studies which have utilized non-omic data as well, mainly relying on clinical assessments and symptomatology, in the identification of schizophrenia patients. These assessments typically involve structured interviews and questionnaires, such as the Positive and Negative Syndrome Scale (PANSS) or the Scale for the Assessment of Negative Symptoms (SANS). Researchers have examined the association between specific symptoms, their severity, and the diagnosis of schizophrenia, aiming to identify reliable indicators and patterns that can help in the early detection of schizophrenia.

In [2] authors aimed to investigate the relative importance of genetic and demographic factors in the prediction and early diagnosis of schizophrenia. They used data containing cognitive assessments, and physical measurements, provided blood samples, and answered touchscreen questions. The researchers identified specific symptom clusters, such as hallucinations, and delusions, which exhibited a high diagnostic accuracy for schizophrenia. These findings highlight the potential of non-omic data in providing insights for the identification of schizophrenia patients.

Then there are neuroimaging techniques, such as magnetic resonance imaging (MRI), have been utilized to investigate structural brain abnormalities associated with schizophrenia. Studies leveraging non-omic neuroimaging data have focused on identifying specific brain regions or connectivity patterns that may serve as potential biomarkers for schizophrenia. Researchers have examined these regions to better understand the neurobiological basis of the disorder.

Event-related potentials help examine and diagnose schizophrenia. In the study [1], authors addressed a more efficient method for schizophrenia diagnosis which could be applied to event-related potentials (ERPs) data. This study provides machine learning techniques applied to event-related potential (ERP) data for the diagnosis of schizophrenia.

Then researchers have employed various neuropsychological assessments to evaluate cognitive functions such as attention, memory, executive functioning, and social cognition in individuals with schizophrenia. By examining cognitive performance profiles, researchers aim to identify specific deficits that could differentiate schizophrenia patients from healthy individuals. These findings highlight the non-omic cognitive assessments in aiding the identification of schizophrenia patients.

So, existing studies utilizing non-omic data sources have displayed the potential of clinical assessments, symptomatology evaluations, neuroimaging techniques, event-related potentials, and cognitive assessments in aiding the identification of schizophrenia patients. However, there are limitations to relying solely on non-omic data, as they may lack the comprehensive coverage provided by omic data.

As previously stated, there are numerous data types and machine learning tools that can be used to create machine learning models. Each approach has its own features, benefits, and drawbacks. By contrasting several ML algorithms, we intend to determine how omic data such as Gene expression RNAseq, DNA methylation, and microRNA seq contribute to the prediction of schizophrenia in our study.

## Performance analysis

Machine learning offers a variety of techniques for evaluating the effectiveness of the model. Therefore, several research employed various techniques, including Area Under the ROC Curve, sensitivity, specificity, accuracy, and confusion matrix. Some studies used entirely unique and targeted parameters in addition to these machine-learning techniques to determine the success of their models. According to earlier research, these investigations frequently used sensitivity, specificity, and accuracy measurements.

## Available databases

The data repositories for schizophrenia prediction.

1. Schizophrenia Data and Software Tool (SDST)

* It is developed by the Psychiatric Genomics Consortium (PGC).
* Provides a collection of genomic, transcriptomic, and epigenomic data related to schizophrenia.
* It includes data from various omic platforms, such as genome-wide association studies (GWAS), gene expression microarrays, and DNA methylation arrays.

2. Common Mind Consortium (CMC)

* It consists of data related to schizophrenia and bipolar disorder.
* Provides publicly available dataset containing comprehensive genomic, transcriptomic, and epigenomic data from post-mortem brain tissue samples of individuals with schizophrenia and healthy controls.

3. GEO (Gene Expression Omnibus)

* GEO is a public repository hosted by the National Center for Biotechnology Information (NCBI).
* It contains a vast collection of gene expression data, including datasets related to schizophrenia and healthy controls.

4. UK Biobank

* The UK Biobank is a biomedical database that contains extensive health-related information which includes genomic data, from individuals.
* It helps to explore the dataset to omic features associated with schizophrenia.

These repositories provide valuable resources in utilizing omic data for machine learning-based prediction models in schizophrenia research.

# METHODOLOGY AND RESEARCH PLAN

## Methodology in brief

## Detailed methodology

### Schizophrenia data selection

From the available datasets, we choose GEO (Gene Expression Omnibus) data repository hosted by the National Center for Biotechnology Information (NCBI) for our research.

The National Centre for Biotechnology Information (NCBI) is a renowned national resource for molecular biology information that aims to facilitate a better understanding of the molecular and genetic processes that control health and disease through the development of new information technologies. It serves as a critical repository for genomic data, providing researchers, clinicians, and the public with access to an enormous collection of molecular and genomic data.

One of the most important resources offered by NCBI is the Gene Expression Omnibus (GEO) database, which is an international public repository for gene expression and functional genomics data sets. GEO is an open platform where researchers can submit, access, and analyze functional genomics data sets from a wide range of experiments. GEO provides support for Minimum Information About a Microarray Experiment (MIAME)-compliant data submissions and various tools that enable researchers' querying and downloading of experiments and gene expression profiles to analyze gene expression patterns and identify potential biomarkers or therapeutic targets. It accepts array- and sequence-based data, for functional genomics research.

We decided to use the following data in this study by considering the number of samples.

1. Gene expression RNAseq
2. DNA methylation
3. microRNAseq

The following are the GEO accession number and brief details for the selected datasets.

GSE202537

The title of the dataset is "Diurnal alterations in gene expression across striatal subregions in psychosis". Psychosis is a defining feature of schizophrenia and highly prevalent in bipolar disorder. The experiment type conducted was Expression profiling by high throughput sequencing. This study was conducted at the University of Pittsburgh, and the researchers used RNA sequencing to investigate diurnal alterations in gene expression in human postmortem striatal subregions (NAc, caudate, and putamen) in subjects with psychosis compared to unaffected subjects. The dataset contains a total of 215 samples.

GSE84727

The title of the dataset is "An integrated genetic-epigenetic analysis of schizophrenia: Evidence for co-localization of genetic associations and differential DNA methylation". The experiment type is "Methylation profiling by genome tiling array". The study was conducted at the University of Exeter Medical School, and the researchers used genome tiling arrays to investigate differential DNA methylation patterns in whole blood-derived DNA samples from individuals with schizophrenia and controls. The dataset contains a total of 847 samples, including 414 schizophrenia cases and 433 controls.

GSE223043

The title of the dataset is "MiRNA Differences Related to Treatment Resistant Schizophrenia". The experiment type is Non-coding RNA profiling by high throughput sequencing. The study was conducted at the Health Research Institute of Santiago de Compostela (IDIS), and the researchers used high-throughput sequencing to identify microRNA (miRNA) differences associated with treatment-resistant schizophrenia (TRS). The dataset contains a total of 141 blood samples from schizophrenia patients, including those who responded to medication and those with TRS. The researchers compared the blood miRNAs of schizophrenia patients who responded to medication and those with TRS, thus obtaining a 16-miRNA TRS profile.

### Data preprocessing

It is required to undertake pre-processing procedures since raw data are susceptible to noise, corruption, missing, and inconsistent data. These processes are carried out using classification, clustering, association, and many other pre-processing techniques that are accessible. It is important to increase the dataset's quality because poor data can largely reduce accuracy and result in erroneous predictions. Therefore, the best way to handle such issues is through data pre-processing. With the use of cleaning, integration, transformation, and reduction techniques, knowledge may be extracted from the data collection much more easily.

As information is gathered from many sources and from a real-world application, there is always a problem with data that is missing and large discrepancies in the variety of data. As a result, the strategy of data augmentation produces data for machine learning models. To enhance the effectiveness of the machine learning model and reduce reliance on training data. [11]

### Feature selection

The method of feature selection involves choosing the dataset's most pertinent properties for a particular modeling issue. There are three different kinds of feature selection techniques: filter techniques, wrapper techniques, and embedding techniques. In our model, we intend to contrast feature importance and mutual information.

* Forward feature selection

The wrapper methods include this. It will start off with no features. The model will then be improved through an iterative approach to include the best features one by one until it reaches a consistent level of accuracy.

* Backward elimination

This approach is iterative as well. The whole features are where it starts. The least important feature will thereafter be eliminated at each iteration of the procedure. This procedure will go on until a consistent performance.

* Correlation matrix-pearson

Pearson Correlation is used to investigate the relationship between characteristics and the target variables. Selection will be made based on the correlation value.

* Chi-squared

The ChiSquare method can be used to identify whether a variable is independent or dependent on another. This is a part of the filter methods.

* Feature importance

The feature priority property of the model is used to determine the relevance of each feature, and it is most employed with tree-based classifiers. Each characteristic is given a score, and the one with the highest score is chosen.

* Mutual information

The mutual information between each feature and the target variable and use it as a criterion to select the most informative features. Features with higher mutual information are considered more relevant for predicting or explaining the target variable, and thus, they are retained, while features with lower mutual information can be discarded.

### Apply machine learning methods

The next step is to choose the best prediction algorithms for our research. Our selection of algorithms spans a wide spectrum when we compare algorithm performance on cancer studies, as seen in the following list:

* SVM

One of the most popular machine learning algorithms, SVM examines the data used in regression and classification analyses. Many biological applications benefit from its ability to do classification, regression, and even outlier detection.

Here, the instances are divided into various classes according to the classification mechanism's target value. In an N 10-dimensional space, SVM looks for the best hyperplane to divide data points into distinct classes. It is a method of supervised learning.

* K-Nearest neighbor

KNN is a supervised learning technique for classifying data. The data is categorized depending on the nearest points' class. This is yet another approach applied to biological research.

* Decision tree

The supervised learning technique of decision trees is used in classification. The unique quality of the models created using this methodology is how closely they resemble human reasoning. By iteratively dividing the data into its features and based on the current classifications, a decision tree is constructed.

* Random forest

Random Forest significantly improves accuracy by combining flexibility with decision tree simplicity. We have used random forest in this instance to contrast its effectiveness with other approaches. Additionally, investigations in biology employ this.

* Naïve Bayes

Based on the Bayes theorem, Naive Bayes is a probabilistic machine learning classifier. Naive Bayes come in three different types, including multinomial, Bernoulli, and Gaussian.

### Compare performance

Performance evaluation is a crucial component of our research because improving schizophrenia prediction models is one of our main goals. In order to double-check our findings, we used a variety of techniques in this study, including the Area under the Receiver Operating Characteristic (ROC) curve and the confusion matrix.

* Area under the ROC curve

A graph called the ROC curve is used to evaluate the effectiveness of machine learning algorithms. This approach uses a curve of probability and varies threshold levels. Here, the ratio of true positives to false positives is displayed. However, this approach would be ineffective. However, if the area under the ROC curve is considered, this ROC curve allows for a superior strategy. It calculates the whole area under the two-dimensional ROC curve, which indicates separability.

It offers a whole evaluation of every classification threshold value. Explaining the most recent absolute value of each data point is preferable. AUC assigns a prediction grade.

It also evaluates the accuracy of the prediction model. A model is superior if it has a higher AUC value.

* Confusion matrix

A confusion matrix contrasts the model's correct prediction with its incorrect prediction. A Confusion matrix's columns represent the actual data, while its rows represent what the machine learning system projected. The number of things we wish to forecast affects how big the confusion matrix is.

* Accuracy

The percentage of accurate predictions for the test data is what is meant by accuracy. By dividing the number of accurate predictions by the total number of predictions, it may be simply determined.

## Timeline

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Semester 6 | | | | Semester 7 | | | | Semester 8 | |
|  | W 1 | W 3-11 | W 12-14 | W 15-16 | W 1 | W 2-4 | W 5-10 | W 10-16 | W 1-3 | W 3-8 |
| Literature review |  |  |  |  |  |  |  |  |  |  |
| Annotated bibliography |  |  |  |  |  |  |  |  |  |  |
| Research proposal |  |  |  |  |  |  |  |  |  |  |
| Data collection |  |  |  |  |  |  |  |  |  |  |
| Data preparation |  |  |  |  |  |  |  |  |  |  |
| Build the model |  |  |  |  |  |  |  |  |  |  |
| Experimenting the model |  |  |  |  |  |  |  |  |  |  |
| Research project report writing |  |  |  |  |  |  |  |  |  |  |
| Research paper writing |  |  |  |  |  |  |  |  |  |  |

# PROGRESS TO DATE

## Literature review

## Database collection

### Phenotype data selection

### Data set selection

## Database preparation

# REFERENCES

[1] E. S. Febles, M. O. Ortega, M. V. Sosa, and H. Sahli, *Machine learning techniques for the diagnosis of schizophrenia based on event related potentials*, 2022. doi:10.1101/2022.03.24.22272090

[2] A. De Rosa *et al.*, “Machine learning algorithm unveils glutamatergic alterations in the post-mortem schizophrenia brain,” *Schizophrenia*, vol. 8, no. 1, 2022. doi:10.1038/s41537-022-00231-1

[3] Y. Liu *et al.*, *Machine learning reduced gene/non-coding RNA features that classify schizophrenia patients accurately and highlight insightful gene clusters*, 2020. doi:10.1101/2020.06.08.20125906

[4] M. Bracher-Smith *et al.*, “Machine learning for prediction of schizophrenia using genetic and demographic factors in the UK Biobank,” *Schizophrenia Research*, vol. 246, pp. 156–164, 2022. doi:10.1016/j.schres.2022.06.006

[5] C. J. Gunasekara *et al.*, “A machine learning case–control classifier for schizophrenia based on DNA methylation in blood,” *Translational Psychiatry*, vol. 11, no. 1, 2021. doi:10.1038/s41398-021-01496-3

[6] S. Góngora Alonso, G. Marques, D. Agarwal, I. De la Torre Díez, and M. Franco-Martín, “Comparison of machine learning algorithms in the prediction of hospitalized patients with schizophrenia,” *Sensors*, vol. 22, no. 7, p. 2517, 2022. doi:10.3390/s22072517

[7] Y. Liu *et al.*, “Expansion of schizophrenia gene network knowledge using machine learning selected signals from dorsolateral prefrontal cortex and amygdala RNA-seq data,” *Frontiers in Psychiatry*, vol. 13, 2022. doi:10.3389/fpsyt.2022.797329

[8] Q. Yang, Q. Xing, Q. Yang, and Y. Gong, “Classification for psychiatric disorders including schizophrenia, bipolar disorder, and major depressive disorder using machine learning,” *Computational and Structural Biotechnology Journal*, vol. 20, pp. 5054–5064, 2022. doi:10.1016/j.csbj.2022.09.014

[9] S. Liu *et al.*, “A machine learning model for predicting patients with major depressive disorder: A study based on Transcriptomic Data,” *Frontiers in Neuroscience*, vol. 16, 2022. doi:10.3389/fnins.2022.949609

[10] M. Montazeri, M. Montazeri, K. Bahaadinbeigy, M. Montazeri, and A. Afraz, “Application of machine learning methods in predicting schizophrenia and bipolar disorders: A systematic review,” *Health Science Reports*, vol. 6, no. 1, 2022. doi:10.1002/hsr2.962

[11] K. Maharana, S. Mondal, and B. Nemade, “A review: Data pre-processing and data augmentation techniques,” *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, 2022. doi:10.1016/j.gltp.2022.04.020

[12] Y. Sun *et al.*, “OMICS in schizophrenia: Current progress and future directions of antipsychotic treatments,” *Journal of Bio-X Research*, vol. 2, no. 4, pp. 145–152, 2019. doi:10.1097/jbr.0000000000000049

[13] A. A. Saleh *et al.*, “Evaluation of mrna expression level of the ATP synthase membrane subunit C locus 1 (ATP5G1) gene in patients with schizophrenia,” *Biochemistry and Biophysics Reports*, vol. 30, p. 101234, 2022. doi:10.1016/j.bbrep.2022.101234

[14] H. Sabaie *et al.*, “Bioinformatics analysis of long non-coding RNA-associated competing endogenous RNA network in schizophrenia,” *Scientific Reports*, vol. 11, no. 1, 2021. doi:10.1038/s41598-021-03993-3

[15] A. Sathyanarayanan *et al.*, “Multi-omics data integration methods and their applications in psychiatric disorders,” *European Neuropsychopharmacology*, vol. 69, pp. 26–46, 2023. doi:10.1016/j.euroneuro.2023.01.001

[16] J. Guo, “Transcription: The epicenter of gene expression,” *Journal of Zhejiang University SCIENCE B*, vol. 15, no. 5, pp. 409–411, 2014. doi:10.1631/jzus.b1400113

[17] I. Berardelli *et al.*, “The importance of suicide risk formulation in schizophrenia,” *Frontiers in Psychiatry*, vol. 12, 2021. doi:10.3389/fpsyt.2021.779684

[18] Y. T. Jo *et al.*, “Diagnosing schizophrenia with network analysis and a machine learning method,” *International Journal of Methods in Psychiatric Research*, vol. 29, no. 1, 2020. doi:10.1002/mpr.1818

[19] S. I. Khan, A. Islam, A. Hossen, T. I. Zahangir, and A. S. Latiful Hoque, “Supporting the treatment of mental diseases using data mining,” *2018 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, 2018. doi:10.1109/iciset.2018.8745591