NAZARBAYEV UNIVERSITY

High-dimensional Data Analytics for Genome-wide Association Studies

by

Zamart Ramazanova

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Science Engineering and Technology

Date of Completion
April, 2026

High-dimensional Data Analytics for Genome-wide Association Studies

by
Zamart Ramazanova

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Science Engineering and Technology

School of Engineering and Digital Sciences
School of Sciences and Humanities
Nazarbayev University

April, 2026

Supervised by
Prof. Daniele Tosi (lead supervisor)
Prof. Bakhyt Matkarimov (co-supervisor)
Prof. Amin Zollanvari (external co-supervisor)
Prof. Sheida Nabavi (external co-supervisor)

Declaration

I, Zamart Ramazanova, declare that the research contained in this thesis, unless otherwise formally indicated within the text, is the author's original work. The thesis has not been previously submitted to this or any other university for a degree and does not incorporate any material already submitted for a degree.

Signature:

Date:

BLANK

Abstract

Genome-wide association studies (GWAS) have become a fundamental base of genetics by allowing systematic exploration of associations between genetic variants and complex human diseases. GWAS have uncovered thousands of genetic loci associated with various human diseases. Despite this success, they face important challenges due to the high dimensionality of genomic data, in which the number of predictors greatly exceeds the number of observations.

This thesis focuses on the study of the complex psychiatric disorder of schizophrenia, which is the most challenging and baffling condition in human genetics. Schizophrenia has a multifactorial etiology due to the additive contributions of many genetic, environmental, and epigenetic factors, as well as developmental processes over the lifetime. Schizophrenia is characterized by disruptions in cognitive processes, perception, and emotional regulation, and it affects approximately 0.32% of the global population. Because there are no objective laboratory tests for diagnosing the disease, reaching a clinical diagnosis is difficult due to the high degree of clinical variability, as well as biological variability, which limits the translation of genetic findings into clinical practice. Although many loci related to schizophrenia have been detected through large-scale genome-wide association studies (GWAS), traditional downstream methods such as polygenic risk scores (PRS) provide an incomplete explanation of the heritable risk of schizophrenia and do not offer reliable predictions at the individual level.

The goal of the thesis is to evaluate whether high-dimensional machine learning (ML) can be used to predict schizophrenia risk in comparison with models of genetic risk. To achieve this goal, a comprehensive analysis of GWAS SNP data from schizophrenia case-control studies was conducted.

The thesis is divided into three parts. The first part of this thesis focuses on the development of high-dimensional data analytics methods to identify genetic variants associated with schizophrenia. The main objective is to design predictive models and algorithms capable of predicting the risk of schizophrenia based on individual single-nucleotide polymorphism (SNP) profiles. This study uses the genetic association information network (GAIN) dataset obtained from the dbGaP repository, which includes genome-wide association data from European-American and African-American ancestries. Machine learning techniques are applied to develop ethnicity- and gender-specific predictive models. The proposed models demonstrate strong predictive

performance and achieve the highest accuracy for schizophrenia prediction using SNP data. These findings suggest the potential of SNP-based models as future clinical tools for early risk assessment.

The second part of the thesis is related to the external validation of associations and prediction models evaluated independently in the GAIN data. The replication, robustness, and generalizability of findings generated via the original GAIN dataset were assessed by using a separate sample from the Molecular Genetics of Schizophrenia (MGS_nonGAIN) study. Two complementary approaches were implemented: external validation, in which the Random Forest models trained on GAIN were directly applied to MGS_nonGAIN without retraining, and full external replication, where models were rebuilt from scratch using the same preprocessing steps, association testing, BH-FDR thresholds, and hyperparameter tuning procedures. In an independent cohort, the transfer models retained predictive capability but were somewhat reduced compared to the internal validation process. For the replication analysis, the random forest algorithm again had the highest classification capability.

The third part of the thesis is devoted to the study of two important classes of regularized regression techniques based on the sparsity assumption for analyzing high-dimensional genetic data. The first technique, known as L_1 regularization or LASSO (Least Absolute Shrinkage and Selection Operator) has received attention in statistical learning because it can simultaneously perform coefficient shrinkage and feature selection. The second method is Group LASSO which uses an L_1 penalty too, but provides a framework for selecting predictors based on a grouping structure. Both methods are applied to schizophrenia case-control GWAS datasets to construct high-dimensional classification models. Using the cross-validation technique to determine optimal regularization strength, we evaluate and compare the models in terms of shrinkage behavior, predictive accuracy, and biological interpretability. The comparison of classification performance metrics includes a comparison of how different sparsity structures affect the identification of disease-related genetic variations. Results from following the principles provide an overview of how penalized regression methods can help connect the prediction of statistical events to the interpretation of biological observations in complex disorders.

Acknowledgments

I would like to express my sincere gratitude to my lead supervisor Prof. Daniele Tosi, whose guidance and support during the final stage of my Ph.D. program. Since August 2025, he has provided valuable academic oversight, ensured continuity of supervision, and supported the completion of this thesis and preparation for the defense.

I am deeply grateful to my former lead supervisor, who now acts as my external co-supervisor, Prof. Amin Zollanvari, who supervised the majority of my Ph.D. studies (2019-2025). His expertise in statistical signal processing and machine learning shaped the core of this research. He defined the research topic and guided me throughout the study period, including research direction, methodology development, and experimental design. Coming from a background in physics and mathematics, I entered the field of bioinformatics with no prior experience, and I am especially grateful for Prof. Amin Zollanvari's mentorship in this transition. Under his guidance, I developed my skills in data analysis and machine learning from scratch, including programming in Python and R, and working with tools such as WEKA, Tensor Flow, and PLINK. His courses in machine learning and continuous support played a key role in building my expertise in this area. Prof. Amin Zollanvari's mentorship was defined by dedication, patience, and a constant pursuit of excellence, and I feel truly fortunate to have worked under his supervision. Despite Prof. Zollanvari's moving to another institution (to Utah Valley University, U.S.), he has continued to provide valuable academic support and insightful advice.

I would also like to express my sincere appreciation to my internal co-supervisor Prof. Bakhyt Matkarimov, for his continuous support throughout my Ph.D. studies. He helped me and contributed his expertise in biological and computational data analysis. His advice on the computational aspects of the research and his support in the implementation and interpretation of results were highly valuable. I am especially grateful for the financial support provided through Prof. Bakhyt Matkarimov's research projects, within which I was employed as a Research Assistant. This support was made possible through the grant AP23485899 "Computational methods for collateral mutations analysis" funded by the Ministry of Science and Higher Education of the Republic of Kazakhstan, and the CRP grant #111024CRP2009 "Role of the repair of 5-hydroxymethylcytosine residues in the occurrence of drug-tolerant persister cells in cancer" funded by the Nazarbayev University Collaborative Research Program (NU CRP). This support allowed me to fully dedicate myself to my Ph.D. research.

I would also like to extend my sincere thanks to my external co-supervisor Prof. Sheida Nabavi, for her continuous support throughout my Ph.D. studies. Her expertise in bioinformatics and her valuable advice during the research and publication process were greatly appreciated.

Finally, I would like to express my deepest gratitude to my mother and sister for their constant support, encouragement, and understanding throughout this journey. I am also thankful to my colleagues and friends for their support and companionship during my Ph.D. studies. This journey would not have been possible without their support. I am sincerely grateful to all of them.

Contents

List of Tables

List of Figures

Abbreviations

GWAS	Genome-Wide Association Studies
SNP	Single-Nucleotide Polymorphism
EA-F	European-American Female
EA-M	European-American Male
AA-F	African-American Female
AA-M	African-American Male
WGS	Whole-Genome Sequencing
AMD	Age-related Macular Degeneration
DNA	Deoxyribonucleic Acid
CNVs	Copy Number Variations
ML	Machine Learning
AI	Artificial Intelligence
TAN	Tree-Augmented Naïve Bayes
NB	Naïve Bayes
RF	Random Forests
LR	Logistic Regression
LASSO	Least Absolute Shrinkage and Selection Operator
ROC	Receiver Operating Characteristic Curve
SVM	Support Vector Machine
DNN	Deep Neural Network
MGS	Molecular Genetics of Schizophrenia
PGSs	Polygenic Scores
EEG	Electroencephalogram
CV	Cross-Validation
QC	Quality Control
MAF	Minor Allele Frequency
HWE	Hardy-Weinberg Equilibrium
dbGAP	Database of Genotypes and Phenotypes
GAIN	Genetic Association Information Network
NCBI	National Center for Biotechnology Information
NIH	National Institutes of Health
CNP	Copy Number Polymorphism
CMH	Cochran-Mantel-Haenszel

BH-FDR	Benjamini-Hochberg False Discovery Rate
AUC	Area Under the Curve
QQ	Quantile–Quantile
LD	Linkage Disequilibrium
CPE	Cross-population Evaluation
HLA	Human Leukocyte Antigen
IBS	Identity-By-State

Chapter 1

Introduction

1.1 Background and Motivation

Genome-wide association studies are a relatively new way for researchers to identify genes involved in human disease. The aim of genome-wide association studies (GWAS) is to identify and characterize associations between single-nucleotide polymorphisms (SNPs) and allele frequencies that vary systematically across a range of phenotypes associated with disease progression. The subsequent identification of SNPs associated with these traits can provide novel insights into the biological mechanisms underlying these phenotypes. Technological advances allow us to study the effect of a large number of SNPs spread across the human genome [1–3]. The advent of GWAS has changed the manner in which researchers study complex traits such as schizophrenia, autism, bipolar disorder, and type-2 diabetes etc., because it provides researchers with a method to systematically identify the genetic variants associated with disease risk across the entire genome.

Mental health conditions are currently one of the most serious global medical and social problems. Psychiatric traits linked to categorically defined mental disorders are heritable and occur to varying degrees among the general population. Many psychiatric disorders manifest themselves in different ways, for example, schizophrenia, depression, developmental disorders, dementia, intellectual disability, and bipolar disorder, including autism. Psychiatric symptoms tend to be hereditary and are associated with categorically defined mental disorders, and are present to varying degrees. GWAS have significantly increased understanding of the basics of psychiatric and other disorders [2, 4–9]. Although the results of genome-wide association studies provide insight into the genetic basis for complex traits, the challenge still exists to clinically translate these genetic findings into clinical practice. This remains an even more difficult issue for complex psychiatric disorders such as schizophrenia.

Schizophrenia is a complex psychiatric disorder that has been extensively studied in human genetics. Schizophrenia is estimated to affect 24 million or one out of every three hundred people (0.32%) globally, as per the current data provided by the World Health Organization, and among adults with the rate of 1 in 222 people (0.45%) [10–13]. This disorder is characterized by disturbances in cognitive, perceptual, and emotional regulatory functions, and diagnoses are primarily based on clinical evaluations rather than objective laboratory tests. Due to its heterogeneous nature and the considerable

impact of genetic and environmental influences, schizophrenia presents a significant challenge and opportunity for current research [14].

Recently, genome-wide analysis has paved the way for developing machine learning (ML) techniques suited to analyze high-dimensional genomic data. GWAS data is a promising approach for improving disease prediction and understanding the mechanisms behind complex genetic diseases like schizophrenia [15]. High-dimensional methods can analyze genomic data using machine learning. There are several advantages to using machine learning methods for this purpose, including the ability to model nonlinear relationships, handle large amounts of predictor data, and discover complex patterns by integrating data from GWAS. This leads to better prediction of disease incidence and a deeper understanding of disease through genetic variability.

1.1.1 Research gap

Genome-wide association studies (GWAS) have identified many genetic loci for schizophrenia, in particular 108 genetic loci associated with schizophrenia have been observed in these research studies [16, 17]. Many various genetic loci associated with schizophrenia gave reported in studies [16, 18–28]. However, due to the small effect size of the majority of the discovered SNPs associated with schizophrenia, they do not help to clarify how schizophrenia develops. As an example, the odds ratios for the SNPs identified as being associated with schizophrenia are typically in the range of 1.10 to 1.20 [29]. Single-variant association analysis, such as polygenic risk scores, will likely not accurately predict disease when using the variants independently, as the number of variants is so large.

Furthermore, another challenge with the already published predictive models is that no previous studies have evaluated multivariate SNP-based predictive models in a systematic manner, across different population groups. Due to genetic heterogeneity, and potential sex-specific effects, genetic markers and predictive models will differ by ancestry group, but a large percentage of the studies performed to date have used combined populations when conducting analyses without directly considering these differences and how it influences the predictive performance. Along with these challenges, GWAS datasets are typically high dimensional, where the number of predictive genetic variables considerably exceeds the number of observations, which presents significant additional statistical and computational problems. This Ph.D. project aims to bridge the gap between mathematical-statistical advances and multivariate genetic analysis

1.2 The present study

1.2.1 Hypotheses & aims

The research hypothesis is that multivariate machine learning models can combine weak SNP signals to identify schizophrenia-related loci and enhance prediction accuracy, especially when accounting for population structure.

This thesis aims to develop and evaluate high-dimensional data analytics methods that integrate multiple genetic variants using machine learning, in order to identify those associated with schizophrenia and to assess their predictive utility based on the SNP profile.

To achieve this goal, the following research objectives are addressed:

1. To create multivariate predictive models of schizophrenia using GWAS SNP data.
2. To identify the SNP superclasses that has the significant predictive power.
3. To assess the predictive capacity of machine learning techniques for classification of schizophrenia cases and controls.
4. To estimate the reliability and generalizability of predictive models through validation against independent datasets.
5. To evaluate the effect of population structure by creating ethnicity- and gender-specific predictive models.

Using high-dimensional genomic data with machine learning algorithms, we can show that we can predict if someone is likely to develop schizophrenia based solely on their SNP profile. If these models are successful, then they will contribute to understanding the genetics of schizophrenia and possibly lead to new ways to clinically identify individuals who are at risk of developing schizophrenia based on their genetics across different populations.

1.2.2 The subject and relevance of the study

The object of the study is machine learning methods. The subject of the study is a classifier development to predict schizophrenia disease in European American (EA) and African American (AA) populations.

The relevance of this research is that, as increasing amounts of multivariate genomic data become available, there is a growing need to develop more sophisticated analytical methods for handling these types of data and to enhance our knowledge of the genetic basis of schizophrenia. As a result, schizophrenia will continue to be one of the most challenging psychiatric disorders to study genetically due to both its multi-gene

nature and its very complicated interplay between genetic and environmental factors. Developing accurate predictive models that rely on genomic data may facilitate a better understanding of the relative risk of developing schizophrenia as well as future strategies for early diagnosis and individualized medicine.

1.2.3 Practical significance of the study

The practical significance of the study lies in the selection, development, and evaluation of machine learning methods that will support clinicians in the early detection and diagnosis of schizophrenia, potentially enabling earlier intervention strategies. These machine learning methods can be used to find associations between genetic markers and schizophrenia. The predictive models based on SNP profiles may provide future prognostic tools for assessing schizophrenia risk across multiple populations. Also, from a practical perspective, the study provides a valuable addition to ongoing effort to develop predictive models that can eventually support precision psychiatry. This research's meaningful aspect is the explicit consideration of population heterogeneity. The study builds ethnicity- and gender-specific predictive models, namely European-American female (EA-F), European-American male (EA-M), African-American female (AA-F), and African-American male (AA-M) subgroups, the study illustrates that genetic risk is not the same in all groups within a population. This approach helps to reduce confounding effects that often arise in GWAS due to population stratification and demographic differences. In practical terms, such stratified modeling may lead to more robust and interpretable predictions within specific subpopulations. It is also consistent with a broader trend in biomedical research toward more representative analyses, particularly given the historical underrepresentation of certain ethnic groups in genetic studies. The developed models in this work provide a methodological foundation for future translational research.

1.2.4 The scientific novelty of the project

The scientific novelty of the project is to use the power of various mathematical-statistical machineries that suit high-dimensional data analytics in analyzing GWA data. Prognostic models developed in this project can serve as clinical tools to assess the risk of the development of psychiatric disorders in individuals based on their genetic variants. Identifying genetic variants is a main step in improving disease diagnosis. This is because it helps us understand the genetic basis of complex disorders. It also allows us to identify individuals at higher risk, supporting earlier detection and execute personalized clinical decision-making. Furthermore, we aim to estimate the degree of shared genetic risks and to perform the computational framework for identifying biological pathways and processes that significantly discriminate phenotypic groups.

This would have the potential to nominate new therapeutic targets that would be used for pharmaceutical treatments and have immediate clinical and public health benefits.

1.3 Thesis outline

This thesis consists of several chapters, each addressing a distinct theoretical, methodological and analytical aspect of schizophrenia prediction based on the GWAS data and machine learning approaches.

Chapter 1 introduces the research context and provides the general background of the study, outlines the research motivation, identifies the existing research gaps in schizophrenia prediction methods, and then formulates the hypothesis and objectives. It also highlights the significance and novelty of the work.

Chapter 2 presents the theoretical background and literature review. It introduces genome-wide association studies (GWAS), single-nucleotide polymorphisms, and schizophrenia as a complex disorder. In addition, it reviews machine learning techniques relevant to this study, including classification methods, feature selection approaches in GWAS, cross-validation strategies, and bioinformatics tools such as PLINK.

Chapter 3 outlines the methodology used in this research. It describes the datasets, particularly the GAIN schizophrenia dataset, and details the data preprocessing steps, statistical analyses, and the overall analytical pipeline used for model development.

Chapter 4 focuses on external validation and independent replication analyses using the Molecular Genetics of Schizophrenia (MGS) dataset. It presents both the direct application of models trained on the GAIN dataset and the replication approach, in which models are retrained on the independent dataset to assess robustness and generalizability.

Chapter 5 presents the LASSO and Group LASSO classification analyses applied to the GAIN schizophrenia data. This chapter describes data preparation, model construction, hyperparameter tuning using 5-fold cross-validation, and feature selection. It also evaluates the predictive performance and interpretability of these sparse regression methods.

Chapter 6 presents the results and discussion of the study. It includes prediction results for the GAIN dataset, outcomes of external validation and replication analyses using the MGS_nonGAIN dataset, and results of the LASSO and Group LASSO models. This chapter also discusses the findings in the context of existing research.

Finally, **Chapter 7** summarizes the main conclusions of the thesis. It addresses the research questions, outlines the study's limitations, and suggests directions for future research.

1.4 Contributions

This thesis makes the following contributions:

C1: Development of machine learning algorithms for schizophrenia prediction A number of predictive models have been developed to predict schizophrenia among individuals based on high dimensional SNP data derived from GWAS studies.

C2: Ethnicity-gender subgroup analysis Ethnically and gender stratified models have been developed to account for genetic heterogeneity in the genomics of schizophrenia between different population subgroups as a means to improve prediction accuracy and identify ethnicity and gender specific associations with schizophrenia.

C3: Integration of statistical association test and machine learning methods Statistical association tests for use in GWAS studies have been combined with machine learning algorithms as a means of improving feature selection and providing stable predictive models for schizophrenia.

C4: External validation and replication of predictive models The stability and validity of the predictive models has been established using an independent dataset to evaluate model transferability and external replication.

C5: Evaluation of sparsity-based regression methods for genomic data LASSO and Group LASSO regression techniques for use with GWAS data have been used to assess the effect of using a regularization approach for feature selection to predict schizophrenia on the accuracy and interpretability of the predictive models.

1.5 Publications

The following main peer-reviewed publication related to this thesis:

P1. Ramazanova, Z., Matkarimov, B., Nabavi, S., Crape, B., & Zollanvari, A. SNP-based prediction of schizophrenia using machine learning. *Bioinformatics Advances*, 6(1), vbaf219, 2026. https://doi.org/10.1093/bioadv/vbaf219

Other peer-reviewed publications completed during the PhD program:

P2. Ramazanova, Z., Alikul, A., Begimbetova, D., Taipakova, S., Matkarimov, B. T., & Saparbaev, M. (2025). Study of the Patterns of DNA Methylation in Human Cells Through the Prism of Intra-Strand DNA Symmetry. *International Journal of Molecular Sciences* 26(19), 9504. https://doi.org/10.3390/ijms26199504

P3. Ramazanova, Z., Baiken, Y., Matkarimov, B., Baimagambet, Z., Aituov, B., Myngbay, A., & Urazbaev, A. (2025). Revolutionizing breast cancer detection with artificial intelligence and machine learning breakthroughs in imaging and diagnosis: Literature review. *Engineered Science*, 38, 1859. https://doi.org/10.30919/es1859

P4. Urazbayev, A., Issakhanova, B. A., Baimagambet, Z., Ramazanova, Z., Baiken, Y., & Myngbay, A. (2026). ML in breast cancer IHC: Pilot evaluation on non-ideal slides in Kazakhstan. E*Electronic Journal of General Medicine*, 23(3), em732. https://doi.org/10.29333/ejgm/18520

P5. Ramazanova, Z., Baiken, Y., Matkarimov, B., Urazbayev, A., Myngbay, A., & Aituov, B. (2025). An information technology approach to predict breast cancer using machine learning. *Scientific Journal of Astana IT University,* 24, 36–48. https://doi.org/10.37943/24UTRW4400

P6. Insepov, Z., Ramazanova, Z., Zhakiyev, N., & Tynyshtykbayev, K. (2021). Water droplet motion under the influence of surface acoustic waves (SAW). J*ournal of Physics Communications*, 5(3), 035009. https://doi.org/10.1088/2399-6528/abda13

Individual contribution:

Publication P1 constitutes the principal peer-reviewed publication arising from this thesis and forms the scientific basis for the methodological framework, machine learning prediction results, discussion, and conclusions presented in Chapters 3–7. I was the first author and corresponding author of this publication. My individual contributions included the conception of the study, research design, data acquisition and preprocessing, implementation of the machine learning pipeline, feature selection, statistical analyses, model development, external validation and independent replication analyses, biological interpretation of the identified SNPs, visualization of the results, preparation of the manuscript, revision of the manuscript, and correspondence with the journal throughout the peer-review and publication process.

Publications P2–P6 were conducted during my PhD program but do not included as part of the scientific contributions of this thesis. These publications represent collaborative research in related areas of machine learning, bioinformatics, computational biology, and medical imaging, including DNA methylation analysis, breast cancer prediction, and biomedical applications of artificial intelligence. Although they demonstrate the breadth of my research activities during the PhD program and contributed to the development of my research experience in high-dimensional data analytics and machine learning, they are independent of the schizophrenia GWAS research presented in this dissertation.

BLANK

Chapter 2

Theoretical background and literature review

This chapter provides theoretical background and important concepts for this research study. It provides an overview of the topics of genome-wide association studies (GWAS), single nucleotide polymorphisms (SNPs), about the characteristics of schizophrenia disorder, and the major challenges associated with analyzing high-dimensional genetic data. In addition, this chapter reviews previous GWAS studies of schizophrenia and identifies major findings and limitations of each of the research studies. The first two sections of this chapter will introduce the basic concepts of GWAS and SNPs, which serve as the basis for further analysis.Section 2.3 will discuss the criteria of schizophrenia and symptoms. Section 2.4 will outline important discoveries and conclusions drawn from past research in this area. Sections 2.5–2.7 will present machine learning techniques and methods used in this study and will discuss important methodological elements including cross-validation and feature selection. These sections will provide the necessary background for using machine learning predictive models applied to genomic data. Section 2.8 will present information on the PLINK program, which is used as a tool in the makeup of the GWAS database, to preprocess and manage the data for this study. This chapter provides the supporting foundation for the use of machine learning methodologies to GWAS database explored in subsequent chapters.

2.1 Genome-wide association studies

Genome-wide association studies are a relatively recent approach for researchers to identify genes involved in human disease. Genome-wide association studies (GWAS), also known as whole-genome association studies, compare the genomic differences between case and control groups across the entire genome at high coverage to detect genetic variations by identifying associated markers, specifically single-nucleotide polymorphisms (SNPs). Detecting genetic variants is an important step toward improving the diagnosis. GWAS design is shown in Fig. 2.1, allows to test hundreds of thousands to millions of genetic variants across the genomes of many individuals to identify associations between genotypes and phenotypes has transformed the field of

complex disease genetics over the last ten years [1, 5, 30]. A GWAS aims to identify all single-nucleotide polymorphisms (SNPs) that are relevant to the diseases. Initially, a GWAS identifies the disease or trait to investigate and chooses a suitable study population, such as cases and controls for a disease or a random population sample for a trait. Genotyping is typically conducted using single-nucleotide polymorphism (SNP) arrays with imputation or whole-genome sequencing (WGS). To find genomic regions related to the phenotype of interest at a genome-wide level of significance, association tests are performed. Meta-analysis is often employed to enhance the statistical power to detect these associations. Statistical tests are then performed across all chromosomes to identify variants significantly associated with the phenotype, typically visualized in a Manhattan plot. Typically, the causal variants are not directly genotyped but are inferred through linkage disequilibrium with the genotyped SNPs [30].

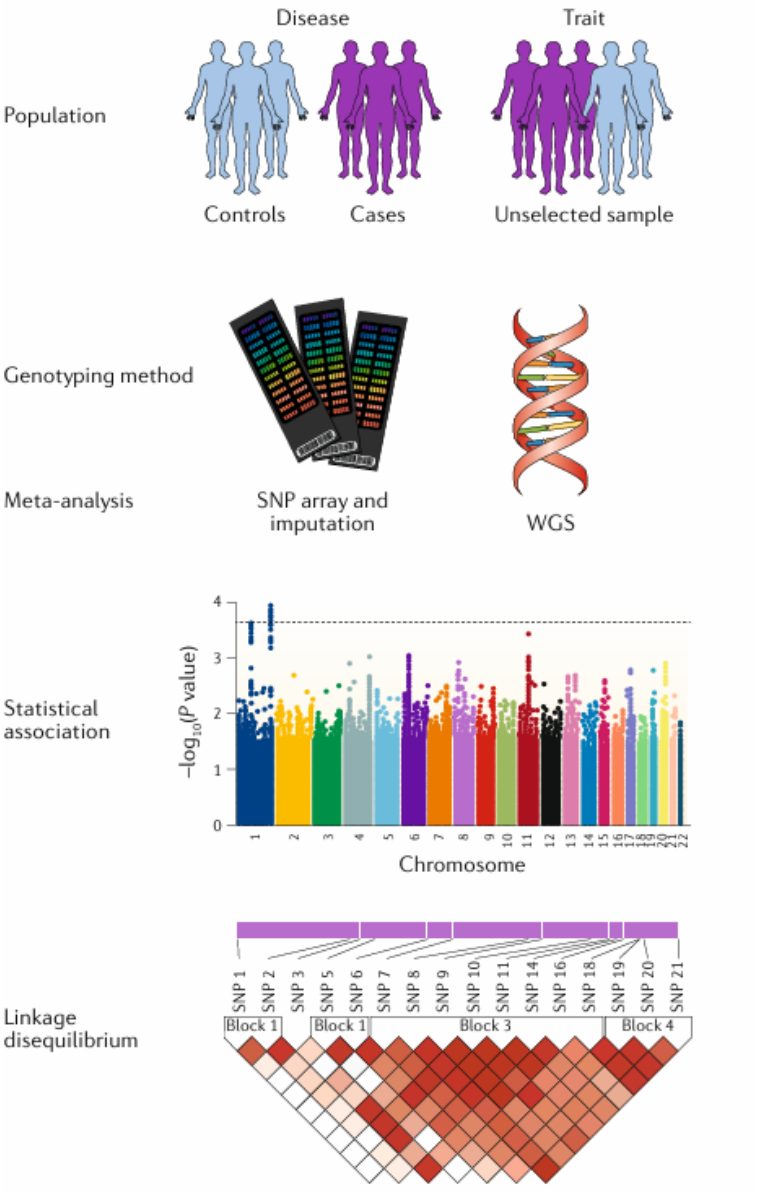

Figure 2.1: GWAS design (from [1]).

The first GWAS for age-related macular degeneration (AMD) was carried out in 2005 that identified two SNPs associated with this disease [31]. But before GWAS were developed, genetic linkage studies were highly effective at identifying single-gene diseases [32, 33]. Linkage studies confirmed that Huntington's disease and Alzheimer's disease are inherited and enabled scientists to identify the genes responsible for these two diseases [34, 35]. However, on the other hand, when analyzing more complex disorders such as schizophrenia, bipolar disorder, or diabetes using linkage studies, researchers have been unable to identify any specific genes associated with those disorders. In 1996, Risch conducted a study to determine if there had been a final limit to genetic research on complex diseases. He suggested that genetic research in complex diseases should be performed using an association approach because results from linkage studies have not been successfully replicated between multiple complex diseases and their associated loci, and based his argument for an association approach, using statistical analyses to support his hypothesis [36]. The results of the human genome project, the creation of biobanks, and the completion of the International HapMap Project have facilitated the progress of GWAS [37].

GWAS is suitable for 'complex' diseases. Genetic variations detected by GWAS can identify individuals who have an increased susceptibility to specific diseases, hence supporting improved patient outcomes through early diagnosis, preventive strategies and more targeted therapeutic interventions. An allele's frequency in the population and its risk associated with complex diseases are important factors to be taken into consideration while designing a genetic study; as such, they will influence the type of technology required for obtaining genetic data and the number of samples required to find statistically significant genetic effects. The potential range of genetic effects can be categorized by their effect size and the allele frequency (see Fig. 2.2).

In Fig. 2.2, Mendelian diseases, such as those shown in the top left circle, have a significant impact on individuals, but mutations of this type are very rare in the population. Similarly, very rare variants with small effects, represented in the bottom left circle, are also present. These factors limit the ability to establish a reliable correlation between phenotype and genotype. GWAS aims to identify many genetic variants, which can be categorized as (i) common variants with high effect sizes (top right circle) and (ii) common alleles that appear to have very little impact on human health (bottom right circle). To detect genetic variants residing in the sample size must be large for two reasons. Firstly, to observe rare variants in a set of individuals, and secondly, to have enough power to estimate low effect sizes.

The use of genome-wide association studies has greatly increased the understanding of the genetics of complex traits by providing a methodical way to identify genotype-phenotype associations across the entire genome. One of the main advantages of GWAS is that they are able to detect many reproducible risk loci that have led to the

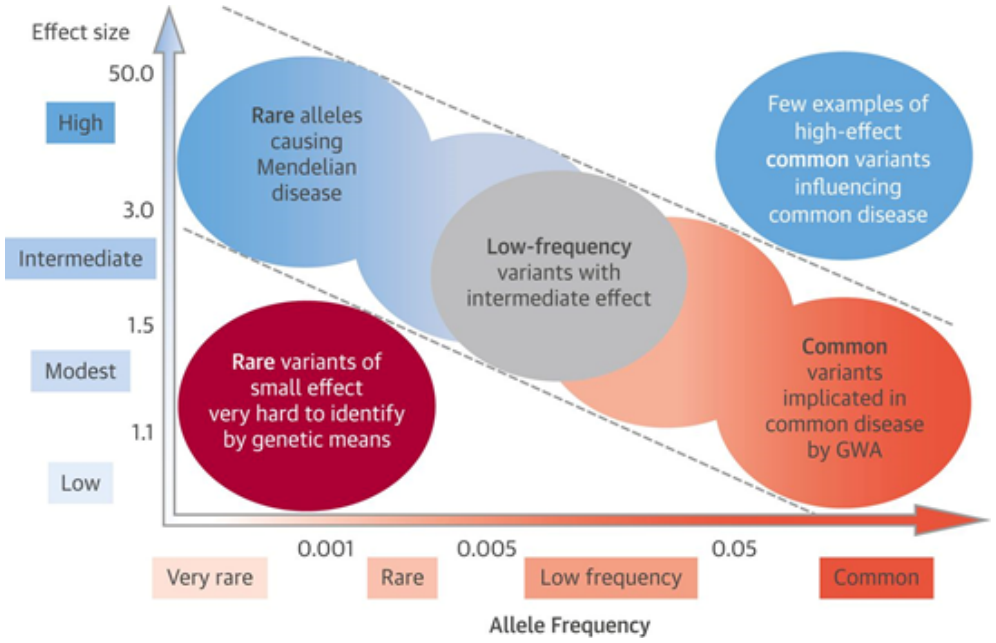

Figure 2.2: A schematic representation of the ability to identify genetic variants based on the risk allele frequency and the strength of the effect size (adapted from [38]).

identification of unknown biological pathways and mechanisms of disease [30, 38]. To date, GWAS have identified lots of SNPs that are associated with psychiatric disorders, including schizophrenia [16, 22, 24], depressive disorder [39, 40], and bipolar disorder [41, 42]. These discoveries have also proven to have translation value because they are used for risk prediction, the development of biomarkers, and identifying potential therapeutic targets. GWAS are organized in a way that allows for them to be scaled, which will allow for an increase in the discovery of risk loci as sample sizes continue to grow with no signs of saturation.

Despite these advantages, GWAS present several important limitations. One limitation is that most of the genetic variants identified through GWAS only account for a small fraction of the heritability observed in traits that are influenced by many genes [43]. Additionally, GWAS results do not provide direct evidence of the biological mechanisms for the genetic variants because of linkages in the genome. This means that extensive laboratory studies will be needed to determine the mechanisms for GWAS results [1, 44]. Furthermore, a large number of estimates must be made when conducting GWAS and therefore researchers use strict significance levels that may reduce their ability to detect genetic variants that facilitate to risk for complex conditions, since there are many fewer studies published reporting small effects than there are studies with large effects [43, 45]. While GWAS are not intended to identify very rare genetic variants, various other factors may diminish their efficacy to identify very rare genetic variants, as well as gene-gene and gene-environment interactions, which may contribute to increased risk of developing complex conditions, when performing GWAS [19, 46, 47]. Additional methodologic limitations associated with GWAS are present, including

potential confounding due to population stratification and the limited clinical predictive value of identified variants, further restrict the applicability of GWAS findings in research and clinical contexts [48–50].

2.2 Single-nucleotide polymorphisms

Single-nucleotide polymorphisms (SNPs) are variations in a single nucleotide occurring at a specific genomic location. These changes or variations are common throughout a person's deoxyribonucleic acid (DNA) and are the most frequent type of genetic variation among people. Figure 2.3 illustrates a SNP, which is a variation in the base at a specific site in the DNA sequence between chromosomes. Each SNP reflects a variation in a single nucleotide, which is one of the fundamental building blocks of DNA. For instance, a SNP may involve the substitution of cytosine (C) with thymine (T) within a particular DNA sequence. SNPs can occur in coding regions of genes, non-coding regions, or in the intergenic regions between genes. While many SNPs have no effect on health or development, some can directly contribute to disease or influence the risk of disease and other health-related traits. SNPs constitute 90 % of genomic variations [51]. These variations can lead to changes in an individual's traits, from susceptibility to diseases to physical characteristics.

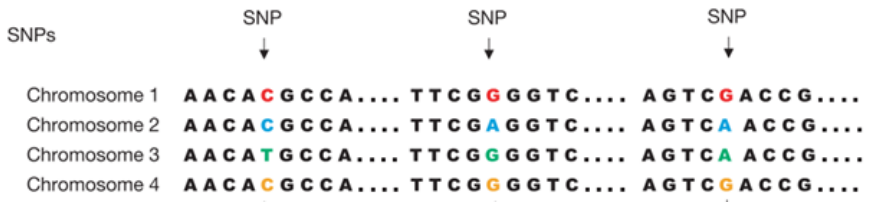

Figure 2.3: SNPs. A brief segment of DNA from four variations of the same chromosomal region in different individuals. While most of the DNA sequence is the same across these chromosomes, there are three bases where differences are observed. Each SNP has two potential alleles; for example, the first SNP in panel a feature the alleles C and T (adapted from [37]).

The location of an SNP can determine its various applications in medicine and research. Coding SNPs associated with a single-gene disorder can be used to predict whether a person will develop the disorder they represent and can be used in approaches to gene therapy. Additionally, SNPs may be useful when studying how a variation of an SNP produces a change in the amino acid sequence of proteins that have a role in drug metabolism [52]. Also, the great numbers of SNPs that have been identified indicate that they are important. Also, because of their relatively low mutation rate, SNPs are ideal markers for use in genomics studies, considering that they demonstrate the historical (evolutionary) development of the human genome [53].

There is currently a high amount of research activity directed towards the identification of profiles of SNPs associated with diseases; as a result, genome-wide studies have improved the ability to identify variation in genetic material as well as developing ways for using this information to identify certain diseases, to prevent illness as well as treat illnesses, and has created additional research opportunities [54].

2.3 Schizophrenia disorder

Schizophrenia is considered to be a heritable mental illness, resulting from a complicated interaction between the individual's genetic predisposition and the environment in which they live [55]. Positive and negative symptoms characterize schizophrenia. The positive symptoms include the presence of hallucinations and delusions, whereas the negative symptoms are the absence of emotions and interests and manifest as disorganized thinking and behavior [55]. People with schizophrenia have an average lifetime risk of developing the disorder of approximately 1%. Schizophrenia is associated with a high level of morbidity and mortality and carries a considerable cost both to the individual and to society [16, 56]. Schizophrenia has been known to have a genetic basis for a long time, primarily because the disorder tends to run in families. However, the best evidence for the relationship between schizophrenia pathogenesis and genetic risk factors is derived from adoption studies and twin studies [57]. Schizophrenia carries a large genetic component, as evidenced by the heritability estimates from twin studies, which range from approximately 83%-87% [58].

Disorganization is made up of two symptoms: loosening of associates and inability to maintain an ongoing line of thinking. However, it may not be true that the three-dimensional model has the best characterization of the various symptoms of this disease [59]. It is also possible that it does not accurately reflect the patterns of the co-occurrence of the symptoms that we have seen in patients [60, 61].

Cognitive deficits are one of the most prominent symptoms associated with schizophrenia. Symptoms can include difficulty sustaining attention or concentration, executive dysfunction, and memory impairment [62]. On average, people with schizophrenia score lower than healthy controls on cognitive performance tests; however, because cognitive performance varies widely among people with schizophrenia, they are more heterogeneously distributed in their performance than people without the disorder. In one of the published studies of cognitive impairments associated with schizophrenia, 2 of the top-scoring individuals on a specific cognitive ability measure were people with schizophrenia [63]. Overall, the individual and societal consequences of these disorders are very significant.

2.4 Genome-wide association study of schizophrenia

Potentially utilizing GWAS to identify possible genetic variants for psychiatric and other disorders has made significant progress in the last few years [2, 5–9]. Assessing GWAS has confirmed that schizophrenia has a polygenic basis, i.e., numerous gene risk variations are typically seen in the form of SNPs [64]. So far, 108 genetic loci associated with schizophrenia have been observed with a sample of 36,989 patients and 113,075 controls [16]. Pardinas at al. identified over 250 additional loci involved in the development of schizophrenia [22]. This work [65] applied a combined sample of 409 families with schizophrenia (263 European-American and 146 African-American populations). The families in this study needed to have an affected proband with this trait and multiple siblings of the proband with schizophrenia or a schizophrenic disease. Associations were found for the previously listed chromosomal locations via nonparametric multipoint linkage analysis of all families [65]. A new location associated with schizophrenia was found in this study [66]. The loci are located nearby in the pseudo autosomal area, and adjacent to the colony-stimulating factor receptor 2 alpha (CSF2RA) gene which is responsible for encoding a protein. According to some GWA studies, differences in the copy number variations (CNVs) of certain chromosomes are associated with a higher degree of risk in developing certain forms of schizophrenia and other psychiatric disorders [67, 68]. A two-phase GWA study was performed in [67] to study differences between CNVs that were present in both phases of the analysis. They reported a nominal association between CNVs at three specific locations: 1q21.1, 15q11.2 and 15q13.3 and the developing schizophrenia. As candidate genes ZNF804A, MYO18B and NOTCH for schizophrenia were reported [21, 27], consistent with the finding of an O'Donovan et al. study that showed evidence for an association of ZNF804A with schizophrenia in a GWAS [23]. In one year, O'Donovan et al. conducted a meta-analysis that resulted in the identification of fibroblast growth factor receptor 2 (FGFR2) as a susceptibility gene for schizophrenia, located on chromosome 10q25-q26 [69]. The study was conducted in a Norwegian discovery sample with 201 patients and 305 controls, and subsequently conducted focused replication analyses in a European sample with 2,663 patients and 13,780 controls and found that PLAA, ACSM1 and ANK3 were also associated with schizophrenia as result of GWAS [70]. Many studies show that schizophrenia is not a simple recessive disorder. Additionally, healthy individuals can carry silent genes that make them more susceptible to schizophrenia [71]. Therefore, similar to other complex genetic disorders, there is widespread agreement that numerous different genetic variants interact to create varying levels of risk for developing schizophrenia, with each variant likely having a small effect size [72].

2.5 Machine learning techniques

The mechanisms underlying the cognitive impairment associated with schizophrenia remain poorly understood. This is due mainly to the heterogeneous nature of the deficits, as well as to the greater complexity and heterogeneity of the overall psychopathology. The area of research also involves the collection of an enormous amount of many dimensional data, which consists of behavioral performance, social factors and environmental factors, complex neuronal pathways that are continuously interacting and cannot be easily analyzed using classical statistical methodologies [73]. However, with the rapidly increasing capabilities of computers, machine learning techniques have become increasingly adept at addressing these methodological challenges [74]. Consequently, there has been a dramatic increase in the use of machine learning (ML) within the field of schizophrenia research, particularly in neuroimaging [75]. Additionally, GWAS that incorporate machine learning methodologies [76] and those that utilize different study designs (e.g., case-control, family study, hypothesis-driven design, intervention) [77] and newer methodologies offers solution to assistance with increased statistical power from schizophrenia associated trait research.

In this thesis, we will use ML to build SNP-based predictive models of schizophrenia that are specific to the ethnicity-gender population. In this thesis, supervised learning methods will be discussed. The next subsection introduces machine learning (ML), defines relevant core concepts, discusses the different classification techniques that were used throughout this study, and provides a brief overview of previously completed studies associated with schizophrenia that utilized ML techniques.

2.5.1 Machine learning concepts

Machine learning (ML), as part of artificial intelligence (AI), focuses on ways to provide computers with the capability to learn through experience rather than through explicit programming [78]. The growth of machine-learning's predictive power, combined with its capability of extracting insight from complex relationships found within large amounts of data, has made machine-learning an increasingly popular technique [79]. Although ML is a part of computer science, it is closely related with math disciplines of statistics, optimization, and information theory. Machine learning is thought to be a way to enhance human cognition through automated algorithmic processing and analysis by identifying important relationships that may not have been observable through traditional manual inspection of data [80]. Different subfields of ML utilize different methods to learn and solve problems as well as various types of data.

The ML methods are classified into three major categories, namely: a. supervised learning (binary or multiclass classification), b. unsupervised learning (dimension reduction), and c. reinforcement (reinforcing learning through immediate reward/penalties).

Reinforcement is used primarily in robotics [75, 81]. Supervised learning uses input data that has been categorized into sets of features and corresponding labels (i.e. the predicted outcome). The data is sometimes split into train and test datasets and the corresponding labels recorded on each of those datasets have been removed. The relationship between the features and their corresponding labels are obtained from the training dataset and will then be validated against the test dataset. On the other hand, when working with an unsupervised method we are trying to find some sort of pattern or structure in a dataset that has not already been categorized with labels.

Analyzing data with machine learning algorithms consists of three major stages: (1) preparing data (data preprocessing and division data into a training and test subset), (2) learning (parameter selection, model turning, and regularization of estimator parameters), and (3) evaluating the model performance (applying parameters to the test set) [75]. In particular, algorithms employ pattern recognition methods to identify patterns in large quantities of data by testing many commonsense assumptions about how the data is organized. Then learn from these assumptions by comparing them and adjusting just one of the factors that make up a model(s). Thus, the process involves many repeats of thoroughly estimating parameters, assessing performance, analyzing errors, and correcting them until the model achieves maximum accuracy [74, 75].

2.5.2 Classification algorithms

This study applies several supervised machine learning classification models, such as Naïve Bayes, Tree-Augmented Naïve Bayes, Random Forest, Logistic Regression with L2 ridge penalty, Lasso (or L1 regularization), and Group Lasso are used to analyze the GWAS data. These models were chosen to examine different methodological approaches, including probabilistic, logistic, sparse, and ensemble-based techniques, particularly applicable for high-dimensional GWAS data in schizophrenia's research.

Naïve Bayes (NB)

Naïve Bayes (NB) is a ML classification algorithm that predicts the categorical response of a data point using probability. The primary concept of NB is to apply Bayes' theorem for classifying new observations based on the probabilities associated with various classes in relation to a set of features [82]. In other words, Naïve Bayes classification assumes that the value of one feature does not influence the value of another feature given the value of the class variable. Typically used for classifying high-dimensional datasets, there is significant interest and work with NB [83]. The probability of a class C_k label given a feature vector $x = (x_1, \ldots, x_n)$ is defined as Bayes' theorem:

$$P(C_k \mid \boldsymbol{x}) = \frac{P(C_k)\, P(\boldsymbol{x} \mid C_k)}{P(\boldsymbol{x})} \tag{2.1}$$

Under the independence assumption, the likelihood term can be factorized as:

$$P(\boldsymbol{x} \mid C_k) = \prod_{i=1}^{n} P(x_i \mid C_k) \tag{2.2}$$

Thus, the classification decision rule is:

$$C^* = \arg\max_{C_k} P(C_k) \prod_{i=1}^{n} P(x_i \mid C_k) \tag{2.3}$$

So, equation 2.3, becomes NB classifier.

The NB algorithm will often work efficiently processing high-dimensional of variables found within genomic datasets. Mostly, NB algorithm is used consistently in the areas of medicine to predict the probability (likelihood) of a person having a specific disease by evaluating the amount and type of their presenting symptoms, in filtering out spam e-mail, for classifying different types of text, predicting the weather, and many other areas as well.

Tree-Augmented Naïve Bayes (TAN)

The Tree-Augmented Naïve Bayes (TAN) model builds on the standard Naïve Bayes model by allowing for inter-feature correlations [82,84]. In this sense, the TAN structure-learning method is a semi-naïve structure-learning method utilized in conjunction with a Bayesian search approach, which is described in [82] and has demonstrated superior effectiveness as a structure-learning algorithm compared to the NB model. The TAN algorithm assumes that all the attributes are conditionally independent given the class attribute. This is done using the class as the only parent of all other attributes. The TAN algorithm then incorporates edges between the attributes to capture the potential dependencies between attribute pairs, given that the class is a parent of all attributes. The TAN also restricts dependencies to one parent other than the class attribute for any given attribute. The NB model assumes that all variables are independent given the class attribute, which can create problems if this assumption does not hold. Studies show that TAN performs better than NB [85].

In a TAN, features will depend on the class variable and no more than one other feature, producing a tree structure. We formulate the joint probability distribution as follows [85]:

$$P(C, X_1, \ldots, X_n) = P(C) \prod_{i=1}^{n} P(X_i \mid C, Pa(X_i)) \tag{2.4}$$

where $Pa(X_i)$ denotes the parent node of feature X_i.

The TAN model is useful for modeling the relationships among genetic variants that may have a combined effect on an individual's risk of developing schizophrenia.

Random Forest (RF)

Random Forest (RF) is an ensemble machine learning algorithm which combines many decision trees to create a better prediction model [86, 87]. The RF algorithm builds multiple decision trees using randomly selected subsets of the training sample and then combines the predictions from each of the individual trees by using majority voting to produce the final output. Using RF improves accuracy and reduces the chances of making predictions that are incorrect. A bootstrap sample of the training dataset is utilized to train each tree in the dataset, with randomly selected features at each split. A majority voting is used for predicting outcomes in classification tasks [85]:

$$\hat{y} = \text{mode}\{h_1(\boldsymbol{x}), h_2(\boldsymbol{x}), \ldots, h_T(\boldsymbol{x})\} \tag{2.5}$$

where $h_t(\boldsymbol{x})$ is the prediction of the t-th decision tree ($t = 1, 2, \ldots, T$); T is the total number of trees in the RF; $\text{mode}(\cdot)$ denotes the most frequent class label among all tree predictions; $\hat{y}$ is the final predicted class label for the input variable $\boldsymbol{x}$.

The Gini importance score is calculated for each individual feature by RF to show how much difference that variable had on differentiating between classes (in the case of classifying trees) as opposed to just using the overall outcome of all the classes together [88]. As such, RF can be used for both classification as well as exploration of each of the individual features [86].

Logistic Regression with L2 ridge penalty

Logistic regression (LR) models can use penalization (or regularization of the coefficients). The effect of these procedures is to shrink the coefficients towards zero, relative to the maximum likelihood estimates. The L_1 and L_2 penalties associated with model fitting are the most common forms of regularization in machine learning. In linear algebra, there are two different types of penalty norms: L_1 and L_2 penalty. Norms provide a measure of how large a vector is based on the distance from the origin to the location of the vector [85]. LR uses the logistic function to represent a binary response variable by estimating the probability of a binary outcome of interest [89]:

$$P(y = 1 \mid \boldsymbol{x}) = \frac{1}{1 + e^{-(\beta^T \boldsymbol{x} + b)}} \tag{2.6}$$

the log-likelihood function of the logistic regression model with L_2 penalty becomes:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N} [y_i \log p_i + (1 - y_i) \log(1 - p_i)] - \lambda \sum_{j=1}^{p} \beta_j^2 \tag{2.7}$$

where $p_i = P(y_i = 1 \mid \boldsymbol{x}_i)$ is the predicted probability that the i-th observation belongs to class 1, $\boldsymbol{\beta}$ is the vector of regression coefficients (excluding the intercept), and λ is a regularization parameter controlling the strength of the penalty term. The term $\sum_{j=1}^{p} \beta_j^2$ represents the squared L_2 norm of the coefficient vector.

This shrinks the effect sizes of covariates (coefficients) towards zero, resulting in better generalizability and stability of models constructed on high-dimensional data generated from GWAS [85].

LASSO (L_1 regularization)

LASSO (Least Absolute Shrinkage and Selection Operator), also known as LL_1 regularization that results in sparsity for the models' coefficients. L_1 regularization can shrink, or set to zero, many of the coefficients in the least-squares estimate, which enables the elimination of uninformative predictors or selection of relevant features. Then the penalized log-likelihood of the LASSO logistic regression model or L_1 regularization is defined:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N} [y_i \log p_i + (1 - y_i) \log(1 - p_i)] - \lambda \sum_{j=1}^{p} |\beta_j| \tag{2.8}$$

where λ is a regularization parameter controlling the strength of the penalty term. The term $\sum_{j=1}^{p} |\beta_j|$ represents the L_1 norm of the coefficient vector.

Unlike LR with L_2 penalty, the LASSO encourages sparsity by shrinking some regression coefficients exactly to zero, thereby performing automatic feature selection. First, the Lasso was introduced by Tibshirani [90]. This is advantageous when modeling GWAS data, since the number of SNPs will usually be considerably greater than the number of subject samples.

Group Lasso

Group Lasso extends LASSO to cases where the predictors can be divided into a priori defined groups. Group LASSO enhances the LASSO framework via variable selection at the level of predetermined groups of variables rather than individual predictor variables. In a GWAS, the groups can correspond to genes, pathways, or genomic regions.

The penalized log-likelihood function is defined as:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N} [y_i \log p_i + (1 - y_i) \log(1 - p_i)] - \lambda \sum_{g=1}^{G} \|\boldsymbol{\beta}_g\|_2 \tag{2.9}$$

where G is the number of predefined groups, $\boldsymbol{\beta}_g$ is the vector of coefficients corresponding to group g, and $\|\boldsymbol{\beta}_g\|_2$ denotes the L_2 norm of group g, defined as $\sqrt{\sum_{j \in g} \beta_j^2}$.

$$\|\boldsymbol{\beta}_g\|_2 = \sqrt{\sum_{j \in g} \beta_j^2}$$

Group LASSO employs the L_2 norm within each group. This is because sparsity is imposed at the group rather than at the individual predictor level. The L_2 norm assesses the total magnitude of the group coefficients without having to constrain individual

coefficients to zero, whereas the outer group-wise summation essentially functions as an L_1-type penalty. Thus, the Group LASSO method retains or removes entire groups of variables. This property is particularly advantageous for GWAS data, as SNPs can be easily grouped by genes or pathways, thereby providing more meaningful, biologically relevant interpretation [85].

2.5.3 Machine learning application in schizophrenia research

ML has been successfully applied in the area of schizophrenia for diagnosis, clinical prognosis, and treatment [74]. For example, the effectiveness of utilizing multivariate pattern recognition analysis on neuroimaging data to separate those suffering from schizophrenia and non-sufferers is demonstrated by a meta-analysis that combined 38 studies, where patterns of sensitivity and specificity for separation were found to be approximately 80% [91]. The development of schizophrenia prediction using human germline genetic materials from the UK Biobank database has been studied in the research [92]. Therein, a range of predictive models has been generated using many machine learning techniques, including distributed random forests, gradient boosting regression trees, XGBoost, GLMs, and stacking techniques that combine the models mentioned above. In terms of performance, the stacked ensemble model performed best and produced an accuracy of 54.5%. In addition, a systematic review [93] demonstrated good potential for diagnosing schizophrenia via machine learning analyses of functional and structural MRI data with accuracy 60-80 %. It was also suggested that with the use of various machine learning methods/tactics, it would be possible to further improve upon current accuracies. In a study [94], researchers used minor alleles from publicly available databases to predict schizophrenia. They found that their predictions were accurate for 3.14 % of patients using a receiver operating characteristic curve (ROC AUC), using 10-fold cross-validation. In another study [95], researchers trained four types of machine learning classifiers (naive Bayes, support vector machine (SVM), random forest and gradient boosting) to classify cases of schizophrenia versus healthy controls; the best-performing classifiers were the random forest with a mean accuracy of 68.6 % and AUC of 60.8%. These studies demonstrate that machine learning techniques may be effective in identifying patients with a diagnosis of schizophrenia based on genetics alone, although more research is needed in this area. In a study [96], it was found that patients who had compromised and spared cognitive functioning were classified correctly $\leq$ 60 %, but that increased to 83 % when using the same sample but looking only at females. Additionally, machine learning algorithms have been shown to be able to predict the outcome of treatment and/or clinical course in patients with schizophrenia across various studies. Specifically, these algorithms have shown to have $>$ 70% (e.g. predicting the outcome of first-episode psychosis) [97] and 85 % (e.g. predicting response to repetitive transcranial magnetic stimulation) [98]

accuracy in predicting outcomes from various types of treatments. Other studies have looked at the potential value of using polygenic scores (PGSs) from genetically related disorders and phenotypes as predictors for schizophrenia risk using both LR and deep neural network (DNN) models [56, 99]. The Molecular Genetics of Schizophrenia (MGS) [56, 100] combined with the Swedish Schizophrenia Case-Control Dataset (SSCCS) [56, 101] were used for training and testing, while the Clinical Antipsychotic Trials for Intervention Effectiveness (CATIE) dataset was evaluated further [56, 102, 103]. DNN models based on the MGS + SSCCS data containing 19 PGS predictors gave an accuracy of 81.3% and AUC of 90.5%, while the CATIE datasets contributed to accuracy (72.1%) and AUC (74.7%) upon model construction. It has been shown that RF model analyses of electroencephalogram (EEG) data can differentiate between patients having schizophrenia and healthy subjects with an accuracy up to 100% [104]. In addition, RF models have been developed with strong predictive accuracy when predicting cognitive subtypes of schizophrenia based on their associated genomic profiles [105].

While ML approaches in schizophrenia research indicate supportive outcomes, careful consideration is needed before implementing them because of problems including: bias in selection of participants subsequently tested, overfitting of data with the model, having subjects who had been previously identified with a similar diagnosis characterized differently in their original diagnosis, a clinical population with varying levels of knowledge about computation and computer-based approaches to assessing patients' functioning, and a lack of sufficient details associated with the presentation of the model and data [74, 75, 106].

2.5.4 Recent advances in deep learning for genomic and multi-omics data

In recent years, deep learning has become an important research direction in computational genomics. While traditional machine learning algorithms such as RF, LR, SVM, and NB remain widely used for structured GWAS datasets, recent studies increasingly employ deep neural networks to analyze large-scale genomic and multi-omics data. These approaches enable learning of complex nonlinear relationships and interactions among different molecular layers.

Recent developments in complex disease studies have highlighted the importance of multi-omics approaches, which integrate different types of omics data, such as genomics (genetic variation including SNPs and DNA sequence), transcriptomics (RNA), epigenomics (like DNA methylation or histone changes), proteomics (proteins), and metabolomics data, that are all used to investigate a more comprehensive understanding of disease mechanisms. Unlike traditional GWAS analyses based solely on SNP array data, multi-omics approaches integrate multiple molecular layers and determinants of

latent biological interactions associated with disease susceptibility. Unlike single-omics studies, multimodal learning approaches are capable of capturing complex biological interactions that may remain undetected in single-omics analyses.

In parallel, multimodal deep learning has emerged as a promising computational framework for integrating heterogeneous omics data. Deep neural networks, including deep learning architectures such as genomic and transcriptomic foundation models HyenaDNA, Geneformer, DNABERT, Enformer and scGPT have demonstrated the ability to learn complex genomic representations directly from DNA sequences and large-scale omics datasets. These foundation models are pre-trained on large-scale genomic or transcriptomic data and can be fine-tuned for downstream tasks such as variant prioritization, gene regulation prediction, cell-type annotation, and disease classification. Their ability to learn transferable biological representations has made them an important direction in computational genomics and precision medicine. These models may facilitate the discovery of novel biological mechanisms underlying schizophrenia and other highly polygenic disorders. Although these models have demonstrated promising performance, their application generally requires large-scale sequencing or multi-omics datasets that differ substantially from the SNP-array GWAS data used in the present study.

Compared with classical machine learning algorithms, deep learning models automatically learn hierarchical feature representations from raw input data and generally achieve the greatest benefit when trained on very large sequencing or multi-omics datasets. In contrast, classical machine learning algorithms such as Random Forest operate directly on structured SNP variables and are better suited to high-dimensional tabular GWAS datasets with relatively limited sample sizes. Moreover, Random Forest provides feature importance measures that facilitate biological interpretation of disease-associated SNPs, whereas deep learning models often require additional explainability techniques to interpret their predictions.

However, the current research focuses on SNP-based GWAS datasets. These datasets contain genotyping information only and do not include additional transcriptomic, epigenomic, proteomic, or other omics modalities required for multimodal deep learning. Furthermore, after ethnicity- and gender-specific stratification, the available sample sizes within each ethnicity–gender subgroup were relatively limited for training large deep neural networks. Under these conditions, deep learning models are more susceptible to overfitting and may exhibit reduced generalization performance because the number of trainable parameters substantially exceeds the amount of available training data. Therefore, considering the characteristics of the available SNP-array GWAS data, the relatively limited subgroup sample sizes, and the need to preserve biological interpretability, random forest was considered more appropriate than deep learning models for the objectives of the present study. Random forest is particularly well

suited for structured high-dimensional SNP-array GWAS data because it effectively captures nonlinear relationships while remaining relatively robust to overfitting through bootstrap aggregation and random feature selection. Furthermore, the algorithm provides measures of feature importance that facilitate the biological interpretation of selected SNPs, which was one of the main objectives of the present study.

2.6 K-fold cross-validation

K-fold cross-validation (CV) is a common resampling procedure for evaluating the performance and generalizability of ML models. *K*-fold CV works by dividing your data into k separate equal-sized subsets (or folds). During each of the k iterations of the CV process, one-fold is used as a validation set, while the other remaining $k-1$ folds are used for training set. This process is repeated k times, ensuring that each data point is used for both training and validation at least once. In this thesis stratified 5-fold cross-validation was utilized in order to keep the ratio of samples within each fold equal between the cases and controls. This is especially important when conducting genetic studies as class imbalance can affect the accuracy of classification models. Additionally, CV can be used to reduce the risk of overfitting and produce a better estimate of the performance of your model using new data (unseen samples). There was both model selection and hyperparametric tuning completed based on the mean results of each of the folds to produce robust predictive models [85].

2.7 Feature selection in GWAS

Feature selection is a crucial step in GWAS, particularly because genomic data is often high-dimensional. This means the number of SNPs can be very large, sometimes reaching hundreds of thousands or even millions. The purpose of applying feature selection is to reduce the dimensionality of a given training dataset by selecting the most relevant features for the phenotype being analyzed. The choice of feature selection is crucial to create a model that generalizes well to unmeasured cohorts. Typically, one of the methods used for selecting which features should be included in the predictive models is to reduce the number of dimensions in the training data by excluding those SNPs that either: a) provide little or no predictive power for the phenotype class, and b) have redundant information to one another [107]. When done effectively, feature selection will improve both the efficiency of the learning process and the accuracy of the predicted values, as well as simplify the complexity of the learned results [108, 109].

In this study, we used feature selection to identify subsets of informative SNPs associated with schizophrenia. Feature selection was done using a statistical association test which evaluated each SNP for an association to case-control status.

2.8 PLINK

PLINK is a widely used open-source software program designed for the analysis of large-scale genotype and phenotypic data. Use of PLINK [110] allows researchers to efficiently perform data management, quality control, and statistical analysis in GWAS data. Consequently, PLINK generally works with two main types of genetic data files: VCF and binary files (.bed, .fam, and .bim).

In the thesis, PLINK 1.9 was used for the preprocessing and analysis of SNP data. Specifically, PLINK was used to conduct quality control (QC) procedures, including the filtering of SNPs and individuals based on missing data, missing allele frequency (MAF), and Hardy-Weinberg Equilibrium (HWE) prior to SNP association analyses with schizophrenia status in order to produce the statistical values necessary for subsequent feature selection. By using PLINK within the analysis workflow, we efficiently processed high-dimensional genomic data while utilizing standardized methods for data preparation before ML modeling [85].

BLANK

Chapter 3

Methodology

This chapter explains of the methodological framework at the core of this thesis to analyze GWAS data and develop predictive models for schizophrenia. It describes the overall analytical pipeline, starting from data preparation to model evaluation.

3.1 Data description

3.1.1 Database of genotypes and phenotype (dbGAP)

The U.S. National Center for Biotechnology Information (NCBI) has developed a public repository known as the Database of Genotypes and Phenotypes (dbGaP) [111], which is a central location for storing, archiving, and sharing research data that studies the relationship between genotype and phenotype. dbGaP was created to meet a need to create a standardized repository to archive large genomic databases, including data from GWAS, sequencing projects, and clinical trials. dbGaP contains two types of information: genotype data and phenotype data. Genotype data consist of variations in DNA sequence such as SNPs while phenotype data consist of measurements of an individual's phenotype (such as health status, physical characteristic, or treatment response). dbGaP houses individual-level data on phenotypes, exposures, genotypes, and sequences, along with their interrelationships. Each study within the database, as well as subsets of information from these studies have a unique identification number. This allows published studies to reference or cite the primary data in a precise manner. A dbGaP offers two types of access for users. Open-access data permits researchers to examine documents related to summary data on specific phenotypic variables, statistical summaries of genetic information, and the genomic positions of published associations. However, accessing individual-level data requires authorized access [112].

3.1.2 Genetic Association Information Network (GAIN) schizophrenia GWAS dataset

The GAIN schizophrenia GWAS data (accession number phs000021.v3.p2) [19, 21, 23, 65] was used in this thesis project which was obtained from NCBI of the database of Genotypes and Phenotypes (dbGap). The GAIN was established to improve the detection of genetic causes of complex diseases through large-scale GWAS. The GAIN

initiative is supported by both the National Institutes of Health (NIH) and industry partners, offering researchers a way to develop genotypic and phenotypic datasets accessible to the scientific community via publicly available repositories such as dbGaP [112]. This dataset is specifically designed to identify genetic loci for genome-wide association studies. The overall goal of the GAIN Schizophrenia GWAS on the dbGaP site is to identify susceptibility genes associated with the development of schizophrenia. The focus of the participant recruitment was based on an individual's ancestry, with participants of European-American (EA) ancestry and African-American (AA) ancestry. The dataset contains both genotype and phenotype data based on interviews conducted to establish diagnostic criteria, epidemiological questionnaires that assess family relationships, and a variety of demographic information obtained during the study. The criteria included adults, 18 years of age or older, diagnosed with schizoaffective disease or schizophrenia that meet DSM-IV criteria. The participants must have been symptomatic for at least six months and must provide informed consent. To meet these criteria, the participant must have a family member or close friend who can confirm the participant's diagnosis. Participants who do not meet the requirements will be excluded from the sample (i.e., they do not communicate sufficiently in English). Participants who are experiencing psychosis based on a substance use disorder or neurological disorder (e.g., seizure disorder) or have significant cognitive deficits, such as a severe cognitive deficit, would also not meet the inclusion criteria. In addition, a mild cognitive deficit ($IQ \geq 55$) would not be included unless there is a clear indication of schizophrenia [56].

The data used in this dataset were collected using the Affymetrix platform and are based on the use of the Genome-Wide Human SNP Array 6.0. The Affymetrix Genome-Wide Human SNP Array 6.0, often referred to as the Affy 6.0, includes a total of 1.8 million genetic markers, comprising over 906,600 SNPs and in excess of 946,000 probes designed to detect copy number variations (CNV). The platform additionally provides analysis tools that effectively integrate copy number and association data using a high-resolution reference map, as well as a copy number polymorphism (CNP) calling method worked out by the Broad Institute [113].

The GAIN schizophrenia dataset includes 906,600 SNPs and 1,983 African-American and 2,710 European-American individuals, further description regarding the initial unprocessed sample characteristics of schizophrenia GAIN (phs000021:phg000013 (EA), phs000021:phg000014 (AA)) is provided in Table 3.1

Table 3.1: The initial unprocessed GAIN data information (adapted from [56]).

Populations	**Individuals**	**SNPs**	**Controls**	**Cases**	**Missing**
African-American	1983	906600	979	953	51
European-American	2710	906600	1442	1215	53

In this study, we separated the data by ethnicity and gender, which helps us to find genetic variants that might be missed in larger studies that mostly include European populations. To avoid the confounding effects of ethnicity and gender on the results of GWAS [114–117], we constructed population-stratified SNP-based predictive models of schizophrenia, that is, the models are specific to EA-F, EA-M, AA-F, and AA-M populations, where "F" is female, and "M" is male populations. Although both ethnicities may possess distinct genetic loci that make them susceptible to developing schizophrenia [118], males are widely recognized to possess increased risk factors associated with schizophrenia [119].

Table 3.2 shows the number of samples in each subpopulation in the data. All subjects selected for analysis were European-American (male and female samples) and African-American (male and female samples).

Table 3.2: The number of individuals in each group in the GAIN dataset (from [56]).

Population	**Individuals**	**Controls**	**Cases**	**Missing**
African-American females	996	609	359	28
African-American males	987	370	594	23
European-American females	1162	774	365	23
European-American males	1548	668	850	30

The GAIN schizophrenia dataset was selected as the primary model-development dataset because it provides publicly accessible, well-characterized case-control GWAS data collected using a common SNP-based platform. The dataset includes participants of both African-American and European-American ancestry, together with information on sex and diagnostic status, allowing the development and evaluation of ethnicity- and gender-specific prediction models. The initial dataset comprised 4,693 individuals and 906,600 SNPs, providing a suitable high-dimensional setting for investigating whether combinations of genetic variants can distinguish schizophrenia cases from controls. Its large sample size, consistent genotyping platform, detailed phenotype information, and diverse population made the GAIN dataset suitable for the objectives of this thesis. In addition, the availability of an independent schizophrenia GWAS dataset enabled subsequent assessment of model transferability through external validation and replication analyses.

3.2 GAIN schizophrenia data analysis

3.2.1 Data preprocessing and statistical analysis

Before model development, we performed preprocessing and statistical analysis to ensure data quality and identify informative genetic variants associated with

schizophrenia. Due to the high dimensionality and complex biological nature of GWAS data, preprocessing is crucial to improve the reliability of subsequent analyses. To eliminate any artifacts from imputation methods before final analyses, we applied the following criteria to exclude all those SNPs with a minor allele frequency < 10%, a genotyping rate < 90%, and a missing data rate < 5%; SNPs with HWE (p-value < 0.0005) were also removed. The PLINK tool [110] was applied for the quality control to exclude low-quality SNPs and samples in accordance with the general recommendations from NCBI. The data pre-processing steps via the PLINK command included next:

1) Individuals were filtered to the separate groups of males and females. The analysis for each group was done separately.

   PLINK commands:

   plink –bfile path/to/files (bed, bim, fam) –filter-males –make-bed –out output_file

   –bfile path/to/files (bed, bim, fam) –filter-females –m–make-bed –out output_file

2) Individuals with missing phenotypes were removed.

   PLINK commands:

   plink –bfile path/to/files (bed, bim, fam) –prune –make-bed –out output file

3) Genetic variants that failed Hardy Weinberg Equilibrium test with a significance threshold of p < 0.0005 were removed.

   PLINK commands:

   plink –bfile path/to/files (bed, bim, fam) –hwe 0.0005 –make-bed –out output file

4) Genetic variants with the genotyping rate less than 90% were removed.

   PLINK commands:

   plink –bfile path/to/files (bed, bim, fam) –geno 0.1 –make-bed –out output file

5) Genetic variants with minor allele frequency (MAF) less than 10% were removed.

   PLINK commands:

   plink –bfile path/to/files (bed, bim, fam) –maf 0.01 –make-bed –out output file

6) Individuals with too much missing genotype data (more than 10%) were removed.

   PLINK commands:

   plink –bfile path/to/files (bed, bim, fam) –mind 0.1 –make-bed – out output file

After completing these quality control steps, we split the dataset into four subpopulations based upon ethnicity and gender which is shown in Table 3.3.

Table 3.3: The characteristics of populations after quality control (adapted from [56]).

Population	Individuals	SNPs
African-American females	968 (359 cases, 609 controls)	871,174
African-American males	964 (594 cases, 370 controls)	873,935
European-American females	1,139 (365 cases, 774 controls)	750,066
European-American males	1,518 (850 cases, 668 controls)	752,739

Further, we split each ethnicity-gender-specific dataset into training (80%) and test (20%) sets using stratified random sampling to preserve the proportions of each subgroup within the entire dataset. The training set was used to create and train classification models, while the test set (comprising 20%) was solely used to evaluate model performance. The number of training and test samples for four ethnicity-gender group is provided in Table 3.4.

Table 3.4: The number of individuals that were used for training and testing purposes (adapted from [56]).

Population	Training set (80%)	Test set (20%)
African-American females (AA-F)	777	191
African-American males (AA-M)	776	188
European-American females (EA-F)	910	229
European-American males (EA-M)	1,209	309

Following quality control, 4,589 individuals remained across the four ethnicity- and gender-specific subgroups, with approximately 750,000–874,000 SNPs retained depending on the subgroup. In general, the resulting dataset represents a high-dimensional classification problem in which hundreds of thousands of SNP variables are analyzed within relatively small ethnicity- and gender-specific subgroups. These characteristics necessitate the use of feature selection, cross-validation, and robust machine learning algorithms are analyzed within relatively small ethnicity- and gender-specific subgroups.

Across all analyses, association analysis at each stage was performed using the Cochran-Mantel-Haenszel (CMH) test [114, 120] using the PLINK. The CMH test is a statistical technique utilized for analyzing the relationship between two categorical variables while also accounting for some additional variable that may potentially affect the association found. In this study the use of the CMH test was beneficial in estimating associations between SNPs and diseases across strata while simultaneously being able

to control for potential differences between the subgroups. After the CMH tests were performed, the resulting p-values were utilized for feature selection purposes.

3.2.2 Model construction and validation

It is highly important to keep training and test datasets separate to properly assess the predictive capacity of a constructed model. After cleaning data, we randomly divided data into training and test data. We used the selected SNPs as the initial list of SNPs to design the models which contained the ethnicity and gender nodes. To correct for possible selection of SNPs, we performed the 5-fold cross-validation procedure [121]. Stratified 5-fold cross-validation (5-CV) was used along with model evaluation and hyperparameter tuning to select a model for each ethnicity-gender training dataset. We developed our own pipeline that used for model construction and a 5-CV is provided in Fig. 3.1 that indicates how this method can be implemented on population specific training data. After preprocessing process and dividing each ethnicity-gender subgroup data into training and test sets, the training dataset was randomly splitted (stratified) into 5 non-overlapping subsets (folds), with all folds having approximately equal sizes. In each round of 5-fold CV, one-fold was held out for a validation set, and the remaining four folds were used for feature selection and model training.

To reduce dimensionality and identify informative SNPs for classification, we applied a feature selection procedure within each gender-ethnicity subgroup. Feature selection was done using the CMH test. This method accounts for stratification and ranks SNPs based on their connection to schizophrenia while also adjusting for group-specific structures. The testing dataset was not used for feature selection purposes and will be used for evaluating the model built from the training dataset. Feature selection and model construction used 4 of the 5-fold and 1-fold was set aside as a validation set. Within each training split, we applied CMH test to the four merged folds to rank SNPs significantly associated with schizophrenia while adjusting for stratification. The features (SNPs) were selected based on a Benjamini-Hochberg False Discovery Rate (BH-FDR) [122] less than thresholds T of $\{0.01, 0.05, 0.1\}$, where T was considered a hyperparameter during the model selection phase. For each iteration, the selected SNPs were used to train a surrogate classifier, and the held-out fold was used to validate model performance. For example, SNPs from folds 2-5 were stored as set X_1 when fold 1 was excluded. This process was repeated using each fold in turn. Thus sets X_1 through X_5 were created, providing an overall assessment of the stability and overlap of selected SNPs across different training sets. SNPs that had a significant association were retained in the training datasets while those without a significant association were removed from them. The area under the curve (AUC) was calculated based on model classifiers that had been built using the training datasets. This process was repeated for each of the 5 subsets and the average AUC was calculated. The total AUC was

also calculated. The maximum CV AUC occurred when T<0.1, thus the final classifier was built using the complete training set (i.e., 776 samples from the AA-M group) and measured only those SNPs with FDR-BH<0.1.

The *"best model"* as determined by AUC on the training data based on cross-validation was the highest AUC model found on the training data by cross-validation. Feature selection approaches were conducted separately on each of the four sub-groups: AA-M, AA-F, EA-M, and EA-F. This gives us the ability to create population-specific models based on the most consistently associated single-nucleotide polymorphisms (SNPs) with schizophrenia within each of sub-groups. All analysis, the entire model training and evaluation were performed through the PLINK tool, R language, Python packages as a Pandas, Scikit-learn module for model classification. The code for reproducing the results is located in our GitHub repository. All code to reproduce the 5-fold CV and model classification is available in the https://github.com/Zamgif/Prediction-Models-of-Schizophrenia/tree/main. More detailed information regarding feature as well model selection is provided in the README.md file in the GitHub repository.

3.2.3 Model Selection

For the purpose of developing ethnicity-gender predictive models for schizophrenia (SCZ), we used four classification approaches, such as Naïve Bayes (NB) [82], Tree-Augmented Naïve Bayes (TAN) [82, 84], Logistic Regression with L_2 ridge penalty (LR) [89], and Random Forest (RF) [86]. The NB and TAN models were chosen for their interpretability and performance on high-dimensional datasets, while the LR and RF models were chosen for their predictive capability [123]. For the LR model, the regularization hyperparameter was varied in $\lambda \in \{0.001, 0.01, 0.1\}$, and the RF model's number of base estimators and maximum depth of decision trees were varied in {10, 100, 1000, 10000} and {2, 5, 10, 20, 50, 100}, respectively [56].

The AUC was used as the performance metric for selecting the best model. Additionally, the accuracy (indicated by "Acc") for each model was also evaluated. The 5-fold CV AUC and Acc values for each model classifier and its hyperparameters are provided in Appendix Tables A.1–A.4 (see Appendix A). These tables indicate the results due to *T=0.1* being the best T value found for each model classifier and training dataset; only the results for *T=0.1* were reported in those tables. Table 3.5 summarizes the AUCs by model classifier for each combination of hyperparameters to present how each model classifier compared to one another with respect to AUC. The results indicate that across all four ethnicity-gender populations, the RF classifier had the highest AUC.

As Table 3.5 shows, the bold line indicates that RF achieved the highest AUC value across all four ethnic-gender groups. At the same time, *T=0.1* was consistently remained the best choice for BH-FDR across all ethnicity-gender populations.

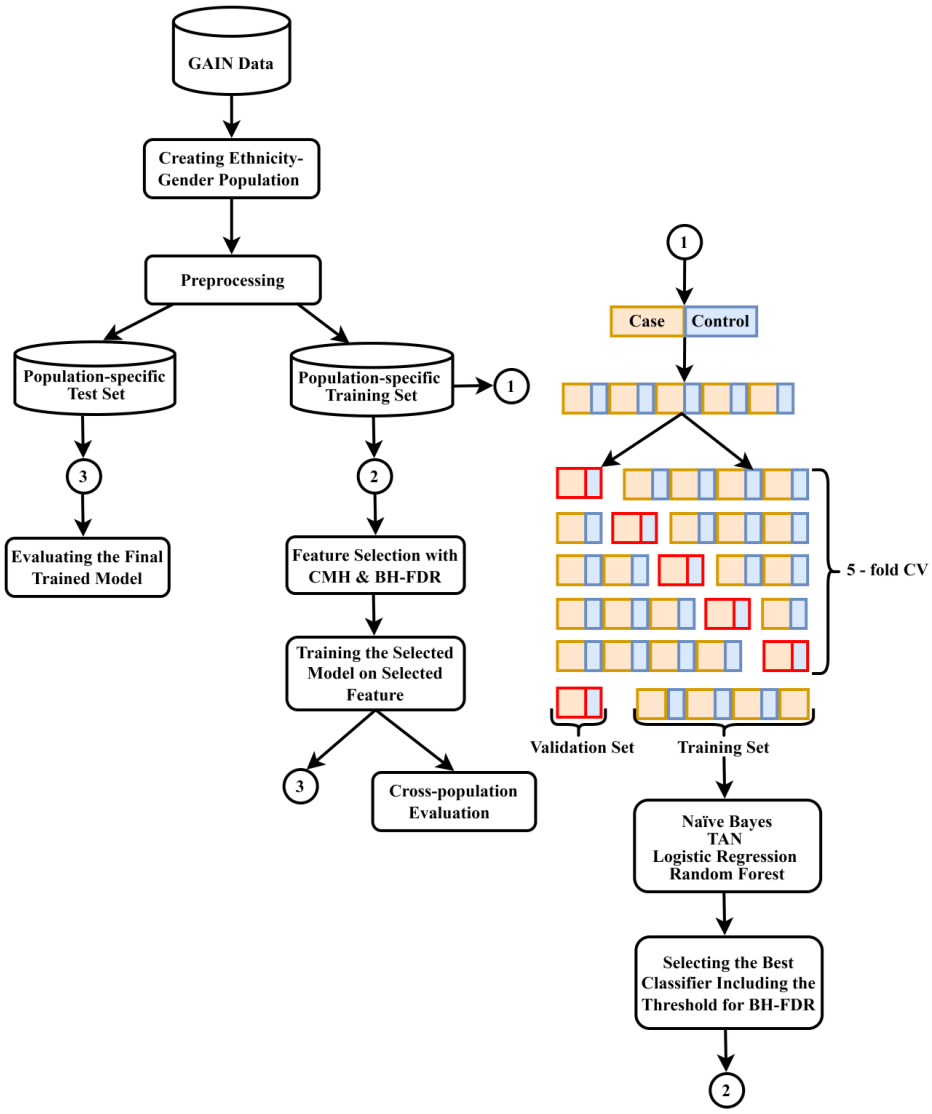

Figure 3.1: A pipeline showing the general steps for data splitting, feature/model selection, and the cross-validation processing for training sets that are specific to each population (adapted from [56]).

Table 3.5: The best 5-CV AUC achieved by each classifier based on all previously discussed hyperparameter values across each ethnic-gender-specific training set (adapted from [56]).

Model	AA-F		AA-M		EA-F		EA-M	
	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
NB	73.3%	73.2%	67.3%	66.7%	79.0%	79.6%	79.7%	71.7%
TAN	72.3%	73.1%	69.5%	69.8%	80.8%	80.7%	81.2%	71.6%
RF	**79.8%**	**78.1%**	**70.3%**	**70.0%**	**89.2%**	**84.1%**	**90.7%**	**80.0%**
LR	71.9%	71.4%	68.6%	69.7%	79.0%	80.7%	81.2%	71.5%

3.2.4 Model training and evaluation

Based on the model selection step which is described in the Methodology Section 3.2.3, across all ethnicity-gender groups, random forest is the preferred model. We trained one last model on each training dataset using the RF hyperparameters values that were identified during the model. In this respect, SNPs specific to each population were established using the CMH test and BH-FDR with a *T=0.1*, which is the 'optimal' value identified from the model selection step (see Section 3.2.3 *Model Selection*). Table 3.6 indicates how many SNPs were selected for each ethnic-gender training set. The RF models constructed from the selected SNPs were then evaluated on the ethnicity-gender test (held-out data) set.

Table 3.6: Selected SNPs number for random forest classifier using *T=0.1* (adapted from [56]).

Population	Number of SNPs
AA-F	7
AA-M	7
EA-F	18
EA-M	37

To mitigate the risk of overfitting, we used 5-fold cross-validation during model training and evaluated model performance on held-out test sets. We also ensured feature selection was performed within the cross-validation loop to avoid information leakage from the test data and ensure a fair estimation of performance.

This chapter describes only the methodological framework, which consists of data preprocessing; statistical analysis; feature and model selection, and development strategies. The proposed pipeline evaluates the effectiveness of various SNP sets in distinguishing schizophrenia cases from controls by incorporating subgroup-specific variability into the predictive analysis.

Several methodological safeguards were incorporated into the proposed pipeline to minimize overfitting and prevent information leakage. Data preprocessing, feature

selection, model selection, and hyperparameter tuning were performed exclusively using the training data within a stratified 5-fold cross-validation framework, while the independent test set remained completely isolated until the final model evaluation. Furthermore, external validation and independent replication analyses using the MGS dataset were conducted to assess the robustness and generalizability of the developed models beyond the original training population. Together, these procedures were designed to obtain unbiased estimates of predictive performance and reduce optimistic performance bias.

The results will be provided in Chapter 6, which consists of analysis and evaluation of the proposed pipeline on test sets, mapping of selected SNPs to genes, prediction analyses based on groupings of different SNPs, and cross-population evaluations.

Chapter 4

External validation and independent replication analyses in the Molecular Genetics of Schizophrenia dataset

This chapter examines how well the predictive models and genetic correlations identified using the GAIN schizophrenia dataset generalize to other independent external datasets. An independent dataset, Molecular Genetics of Schizophrenia (MGS) was used for external validation and replication analysis of this relationship. Two approaches will be used to evaluate the transferability of predictive models trained on the GAIN dataset to the MGS dataset.

The first approach is the predictive models tested on the MGS dataset to evaluate their transferability to a new comparable population. The second approach will be a complete replication analysis from scratch with the MGS data, following the same steps as with the GAIN data to ensure comparable results.

This analysis aims to define the degree of reproducibility of the identified SNP-based signals and predictive models across different datasets. The results of these analyses allow us to assess the stability of the proposed models and their likelihood of being used as valuable tools in schizophrenia genetic studies.

4.1 External validation analysis

4.1.1 Molecular Genetics of Schizophrenia data

The independent external dataset, Molecular Genetics of Schizophrenia - nonGAIN Sample (MGS_nonGAIN) was downloaded from the database of Genotypes and Phenotypes (dbGaP) with an accession number phs000167.v1 [112, 124]. The MGS_nonGAIN dataset provides genotype and phenotype details on 2671 participants who are consenting for General Research Use (GRU) of which there are 1,265 schizophrenia cases, 1,406 controls and 906,600 SNPs. However, it should be noted that the MGS_nonGAIN cohort is different than the GAIN schizophrenia dataset in

terms of the participants making up the cohort and in the clinical spectrum of the cases (no patient overlap). The difference between two GAIN and non-GAIN datasets is that the case and control subjects were genotyped in two phases which are GAIN and NonGAIN genotyping phases. Here approximately half of the EA sample and 95% of the entire AA sample were genotyped within GAIN with the Affymetrix 6.0 platform. The remaining samples were also genotyped with the same platform and referred as nonGAIN sample. The GAIN cohort comprises primarily of individuals who meet DSM-IV criteria to diagnose schizophrenia or closely related psychotic disorders using standardized and harmonized protocols to recruit these individuals. In contrast, although most subjects included in the MGS_nonGAIN dataset meet DSM-IV criteria to diagnose schizophrenia or closely related psychotic disorders, the data collected includes three distinct recruiting efforts (i.e., SGI, MGS1, and MGS2). This approach results in different data from each recruitment effort being used in analyzing outcomes derived from both GAIN and MGS_nonGAIN patients. Control participants in MGS data had to provide a brief screening assessment to exclude them based on any known history of psychiatric illness. To qualify for MGS_nonGAIN, exclusionary criteria were defined for the GAIN study group versus the MGS_nonGAIN. The differences in diagnostic definitions between the DSM-III-R and DSM-IV, the inclusion/exclusion criteria for schizoaffective disorder, and the heterogeneous recruitment for both studies suggest that MGS_nonGAIN had a more heterogeneous population (clinical and etiological) than the group based on GAIN criteria. This heterogeneity and variability of ancestry composition, coupled with location differences in which individuals were recruited from, have the potential to affect baseline allele frequency and to change the structure of linkage disequilibrium. Thus, for example, SNPs that show strong case–control differentiation in the GAIN portion of the sample may be neutral in the MGS_nonGAIN portion, even though they tag the same genomic regions [56].

4.1.2 External evaluation

Here, we conducted an external testing of our findings using the independent MGS dataset. At first, we used the same RF classification model trained on the GAIN dataset and assess how they perform when put through the MGS dataset using the same evaluation metrics (i.e., external validation). A stratification of the MGS_nonGAIN dataset into subpopulations of ethnicity-gender was accomplished using databases from the dbGaP to allow for direct comparison to the GAIN groups. The MGS_nonGAIN sample included mixed-ethnicity (i.e., African-American vs. European-American) populations before stratification. Conversely, the GAIN dataset contained relatively equal populations of AA and EA individuals. The sample sizes of the nonGAIN data after stratification were considerably lower than those of the GAIN dataset where the numbers of cases were 33 and 21 controls for AA-F, 52 cases and 21 controls for AA-M,

356 cases and 685 controls for EA-F, 824 cases and 679 controls EA-M.

All groups were evaluated using the same set of SNPs as identified in GAIN (i.e., 7 SNPs for both AA-F and AA-M, 18 SNPs for EA-F and 37 SNPs for EA-M). So first, the SNPs from the non-GAIN cohort were extracted using the same allele coding that was used for the GAIN dataset. Then, we used the GAIN RF model to predict labels for the non-GAIN subjects and compared those predictions with actual labels in that cohort. This process allows us to evaluate how well the model's performance translates to the new comparable cohort. All SNPs identified by GAIN RF model were also contained within a corresponding subgroup of non-GAIN subjects [56].

The results obtained from external validation are presented in Chapter 6.

4.2 External replication analysis

In the second approach, we built new models from scratch with the MGS_nonGAIN data, following the same steps as with GAIN data, which we discussed in Chapter 2, including preprocessing, feature selection BH-FDR < {0.1, 0.05, 0.01}, and hyperparameter tuning to ensure comparable results. A cross-validation process was performed to train and validate the models, using a stratified 5-fold CV approach to ensure model robustness. We handled a full non-GAIN sample 5-fold CV procedure for testing because we do not want to use the same sample in both training and test data. Here, we only picked up the same SNPs that were part of the classifier, which was developed in the GAIN dataset, and used the same exact classifiers trained on the GAIN dataset to predict schizophrenia in the non-GAIN dataset, because we do not need other new classifier models. Table 4.1 shows a 5-fold CV AUC and accuracies for each classifier model, all the above-mentioned hyperparameter values, and EA-F, EA-M, and AA-F, AA-M samples. Models were developed and validated against multiple independent datasets, specifically the GAIN schizophrenia patient cohort and the MGS_nonGAIN cohort, to facilitate direct comparison of algorithm performance across these datasets. In terms of predictive accuracy on the GAIN dataset, the RF classifier proved to be the highest-performing classifier with respect to the association CMH test and BH-FDR, with a threshold of $T < 0.1$.

Table 4.1: The 5-fold CV prediction results on the training set with different models that are estimated via AUC and accuracy values for AA and EA subjects.

Model	**AA-Females**		**AA-Males**		**EA-Females**		**EA-Males**	
	AUC	**Acc.**	**AUC**	**Acc.**	**AUC**	**Acc.**	**AUC**	**Acc.**
Naïve Bayes	73.1%	73.2%	67.3%	66.7%	79.0%	79.6%	79.2%	71.2%
TAN	72.3%	73.1%	69.5%	69.9%	80.8%	80.7%	80.5%	71.5%
Random Forest	**74.4%**	**75.2%**	**70.0%**	**70.4%**	**86.3%**	**81.2%**	**93.0%**	**83.9%**
Logistic Regression	71.9%	71.4%	68.6%	69.7%	87.8%	85.1%	80.6%	70.8%

Here is in Table 4.1, the highest predictive potential has been shown in the random forest model, which achieved the best performance. All the SNPs that were identified from the GAIN database were also matched across all four subgroups at the genotyping level of all non-GAIN databases. However, overlapping SNPs were retained only in the AA-F and EA-F subgroups at the FDR < 0.1 significance level for association analysis. There is no overlap with either the AA-male or the EA-male subgroups at this threshold. However, there is substantial evidence that the appropriate SNPs are present within the non-GAIN dataset.

The results of the external replication analysis on the test set are detailed in the Results & discussion section of this thesis.

Regarding the external replication, we found that differences between the GAIN and MGS_nonGAIN top SNP lists stem from population-specific allele frequency variation and sampling effects, not from algorithmic limitations. Our algorithm reliably identifies SNPs associated with the disease across independent datasets, confirming its robustness. While genotype overlap exists among all subgroups that overlap only remains under the FDR < 0.10 level for AA-F and EA-F. The changes in statistical significance observed at lower levels of significance are due to changes in allele frequencies and not representative of the lack of these variants.

To clarify this situation, allelic frequencies were calculated separately for both case and control groups in the datasets for AA-M and EA-M. Minor allelic frequencies (MAF) between case and control groups were compared. The difference between case and control frequencies was determined in the GAIN and MGS_nonGAIN datasets. There was a statistically significant difference in allele frequencies between controls and cases, and in some instances. It was determined that the differences are due to random variation across the sample populations, such that certain SNP alleles may appear to be more commonly found in cases than in controls as a consequence of sample-specific effects and not because of true association with disease. A true genetic association with disease would yield a higher MAF in the case samples than in the control samples. However, across our results, one SNP may have high case frequencies in GAIN and low ones in MGS_nonGAIN or vice versa.

To obtain biological validity, we employed rigorous filtering, eliminating SNPs that occur frequently in controls (>1%) and therefore focusing our efforts on SNPs that are consistently enriched in cases across both datasets. This filtering process removes population artifacts and helps identify SNPs that are more likely to be true biological risk factors for schizophrenia. Additionally, our filtering method has the potential to increase the chances that our final SNP set is comprised of variants that have consistent enrichment in cases across populations. As a result, the likelihood increases that our final variant set represents biological contributors to schizophrenia instead of artifacts of the composition of the samples. This shows that our algorithm works reliably both

statistically and biologically.

Also, to visualize the distribution of association statistics, we generated quantile–quantile (QQ) plots for the GAIN EA-M and non-GAIN EA-M datasets for comparison; see Fig. 4.1. The QQ plot for GAIN reveals a marked deviation from the null expectation in the tail of the distribution, indicating an enrichment of low p-values and suggesting the presence of genuine genetic associations.

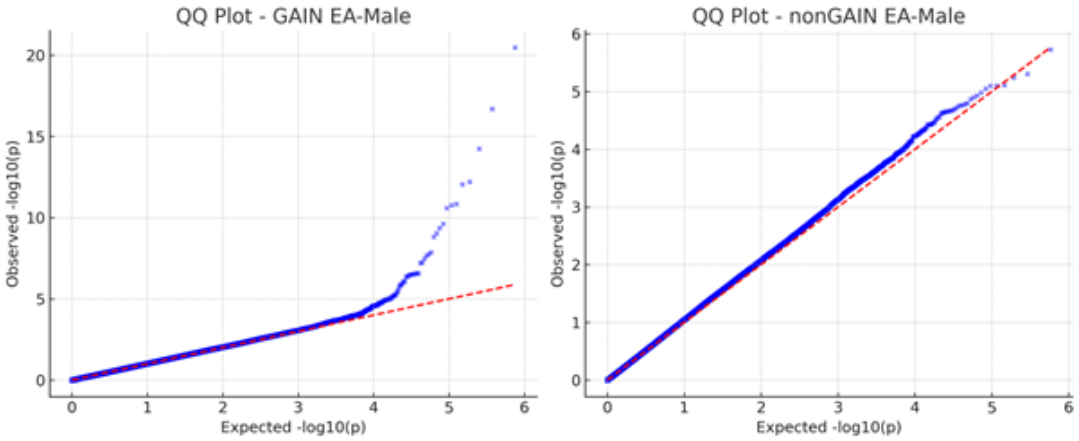

Figure 4.1: QQ plots observed vs. expected p-values for the EA-M subgroup in GAIN and MGS_nonGAIN datasets.

As we can see in Fig. 4.1, the QQ-plot of our dataset is very close to the null distribution of MGS_nonGAIN data. There is a strong apparent contrast in results between the two datasets, showing that the lack of significant SNPs in MGS_nonGAIN is not due to the analysis or methodology but is a true reflection of actual differences between the participant populations, differences in the distributions of allele frequencies, and possibly the strength of linkage disequilibrium in this cohort. The lack of signal in MGS_nonGAIN could be due to differences in proportions of ancestry, sample size, or types of disease represented by the sample, which may produce statistical power that is low enough to produce valid associations or dilute the associated significance of the true association. In GAIN, we found significant signals that indicate our analysis is able to identify meaningful variations in biology in enough quantities to detect them within the dataset, while the failure of replication in MGS_nonGAIN is due to dataset-related issues, not a limit of methodology.

Overall, the results of this study indicate that application of the modeling pipeline (preprocessing, feature selection, classification) is fairly robust across multiple datasets with similar genetic architectures. However, the degree of replication depends critically on the consistency of the underlying data, which in this case varied substantially across certain subgroups. Thus, our model does not merely detect statistical outliers. It isolates variants whose allele frequency shifts between cases and controls are consistent with true disease biology. The presence of strong signals from GAIN but the absence of strong signals from MGS_nonGAIN supports the conclusion that the differences observed

between these two datasets are due to actual biological differences in populations, rather than methodological artifacts.

Chapter 5

Lasso and group lasso classification analysis for GAIN schizophrenia data

In this chapter, we use two sparsity-based regularized regression techniques such as the lasso and group lasso in analyzing high-dimensional genetic data from a GWAS associated with schizophrenia. Whereas traditional statistical methods failed to adequately deal with the "large p, small n" problem that arises when there is a large number of variables (p, or features) and a limited number of observations (n, or samples), penalized regression models can provide an effective framework for performing both feature selection and prediction at the same time. The research focuses on two model-building techniques: the least absolute shrinkage and selection operator (lasso) and group lasso, which use regularization to reduce the model complexity but differ in how they create strictly sparse solutions (sparsity). The lasso creates an L_1 penalty for each of the individual variables (coefficients) with the intention of providing a sparse solution by focusing on individual SNPs, whereas the group lasso as extension of lasso and takes this one step further by grouping variables or SNPs within the analysis to allow for structured feature selection, such as grouping SNPs by genes. The primary purpose of this chapter is to evaluate the effectiveness and compare both approaches for predicting outcomes and with the previous classification models, feature selection behavior, and providing meaningful biological interpretations using to schizophrenia SNP-based data. The theoretical details about two of these regularization techniques, we described in Chapter 2 (see Subsection 2.5.2).

5.1 Data description and model construction

We used the GAIN schizophrenia data as discussed in Chapter 3, particularly in Subsection 3.1.2 *Genetic Association Information Network (GAIN) schizophrenia GWAS dataset* for analysis of the lasso and group lasso. The analysis was conducted separately for four ethnicity-gender-specific subpopulations: AA-F, AA-M, EA-F, and EA-M. After QC, the dataset, given in Table 3.3 (see subsection 3.2.1 Data preprocessing and statistical analysis), had an initial number of SNPs ranging from 750,066 to 873,935

across groups.

High dimensional datasets present a few challenges and issues, one of which is the dimensionality reduction of high dimensional datasets. This is especially true in high dimensional biomedical research since the dataset's number of features can be extensive (e.g., up to millions of variables). For performing this task, both feature selection and feature extraction methods are typically used to reduce the complexity of a dataset, while still preserving its most critical elements. For this reason, the linkage disequilibrium (LD) pruning was applied to the dataset. We performed LD pruning because a large number of SNPs in GWAS datasets are highly correlated due to LD. Therefore, LD pruning was done prior to model fitting to reduce redundancy associated with SNPs and to enhance the stability of penalized regression models. After LD pruning, the reduced number of SNPs is shown in Table 5.1 for each subgroup.

Table 5.1: Population and number of SNPs after LD pruning.

Population	**Individuals**	**SNPs**
AA-F	968 (359 cases, 609 controls)	258,378
AA-M	964 (594 cases, 370 controls)	214,614
EA-F	1,139 (365 cases, 774 controls)	123,342
EA-M	1,518 (850 cases, 668 controls)	74,742

Despite remaining significantly large feature spaces, this reduction is very pronounced and therefore provides lower and more stable outcomes from future penalized modeling than would otherwise be possible. Before model fitting, the genotype data have been placed into a predictor matrix X and binary outcome vector y, with the response variable being the case-control schizophrenia status. Due to SNPs genotypes are represented with categorical features reflecting pairs of alleles, these were converted into numeric representations using Python open-source *scikit-learn* library [125], so they could be used directly by the machine learning algorithms. Using the *scikit-learn* *OneHotEncoder*, the numerical representations of SNP genotypes (e.g., AA, AT, TT) are realized. The majority of the approaches that deal with SNPs involve the counting of minor alleles (ie., AA=0, AT=1, TT=2) for binary encoding. The numerical representation of the biological nature of the DNA variants enabled the models to continue capturing the biological nature of the DNA variants while still permitting the models to be applied to quantitative genotype values. Missing values (if any) were handled during preprocessing, and the predictors were standardized as necessary so that differences in variable scale wouldn't affect the values used in the penalty terms for the regularized models.

The implementation was conducted using a structured pipeline. In the first step necessary Python libraries were imported and data for each sub-groups collected which included both genotype and phenotype. In the second step both the outcome variable

and predictor matrix were established, with schizophrenia case-control status being the output of interest SNP genotypes as predictors. The third step, the data was splitted into training and test sets with 80% and 20% of the dataset size respectively. The final step of the implementation was performing all required preprocessing steps which included encoding genotype data, handling missing genotype data if necessary, and standardizing predictor variables (by subtracting the mean from each value and dividing by the standard deviation of all values).

Two regularized logistic regression frameworks were viewed. The first approach is lasso logistic regression, which has an L_1 penalty on the regression coefficients and reduces many of those coefficients exactly to zero, resulting in sparse models at the SNP level and allowing for embedded feature selection during model training. The second approach is group lasso logistic regression, which uses group-level regularization. Therefore, if one predictor in a particular group is retained, all predictors in that group will also be retained, and if one predictor is removed from that group, all predictors in that group will share the same fate as the removed predictor. The comparison is critical since the two techniques differ in their respective approaches to achieving sparsity. The lasso method examines each SNP separately, lending itself to greater flexibility should the true predictive signal appear to be approximately evenly distributed among several loci. The group lasso method produces an appealing option when developers organize predictors into logically grouped biological clusters. However, it is less flexible than the lasso methodology to the degree the grouping is not representative of the actual underlying disease architecture that governs signal creation across many SNPs.

The lasso logistic regression model was trained on the training dataset. The key tuning hyperparameter in lasso is the regularization strength that is denoted as λ or equivalently by its inverse C=1/λ. When λ=0, the model reduces to a standard logistic regression. As λ increases, the degree of shrinkage becomes stronger, which forces more coefficients equal to 0. If the value of λ is too high, then many of the predictor variables would no longer be included in the final model. Therefore, the choice of tuning hyperparameter is critical for finding an appropriate balance between bias, variance, and predictive power.

The group lasso framework was also developed based on this principle. However, in this case, the penalty-controlled sparsity over defined groups of variables was defined on the basis of SNP phenotypes (groups of individuals with similar traits) instead of SNPs. In both cases, hyperparameters were systematically tuned via cross-validation.

5.2 Hyperparameter tuning using 5-fold cross-validation

To define the optimal regularization strengths, stratified 5-fold cross-validation method was used separately for each ethnicity-gender-specific training set. Training set is splitted into 5 folds of equal size. During each iteration process, four folds were used for training the model and the remaining fold was used as a validation set. After performing this procedure five times, each fold had functioned as the validation set once. The mean of the validation performance across all five folds was used to select and determine a model performance measure. Each regularization parameter λ selected value determined a predetermined grid and were examined. Thus, it covered many regularity penalty options including in $\lambda \in \{10^{-8}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. The purpose of using this range of regularization level is to establish bounds of very little regularization (keeping many SNPs in the model) through to very high levels of regularization (giving very sparse model solutions).

For the lasso model, the inverse regularization parameter C=1/λ was tuned across candidate values. As a result, large values of C represent lower degrees of regularization and weakly penalize any deviation from the regression line. Conversely, small values of C correspond to high degree of regularization and thus, strict penalties on deviations from the regression line. For the group lasso model, the regularization parameter λ and the corresponding best decision threshold were selected. Cross-validation performance is reported in Table 5.2.

Table 5.2: The highest 5-fold CV AUC achieved by each classifier across all aforementioned hyperparameter values for each ethnicity-gender-specific training set.

Model	**AA-F**		**AA-M**		**EA-F**		**EA-M**	
	AUC	**Acc**	**AUC**	**Acc**	**AUC**	**Acc**	**AUC**	**Acc**
Lasso	93.2%	86.1%	93.6%	86.5%	90.0%	85.9%	92.7%	84.7%
Group lasso	88.6%	65.6%	90.4%	67.7%	80.9%	73.9%	78.3%	69.4%

Lasso model shows consistently strong performance with AUC and accuracy values in all four subpopulations. The lasso model implemented in this investigation identified its optimal parameter through use of cross-validation was C=3.16228, which corresponds to $\lambda \approx 0.316$. This value of C is not included as a predefined grid value in the current analysis, rather, it is derived from a logarithmic scale of C's value and the function of C and the location within the logarithmic scale as defined during cross-validation. The outcome demonstrates that a moderate level of regularization produces the best balance between predictive accuracy and model sparsity, which will prevent both underfitting (occurs when λ is very high) and overfitting (occurs when λ is very low).

For the group lasso model, the best predictive performance was achieved at λ=0.1 as measured by the highest cross-validated AUC. Therefore, moderate regularization is optimal for group-based feature selection. However, in comparison with the lasso model, the group lasso is more susceptible to different parameter choices. Hyperparameters that were obtained from performing cross-validation were then applied to the final training dataset for the purposes of building final models. The optimized models were evaluated on independent test sets in order to compare generalization performance to the final models. Chapter 6 presents and discusses those results.

5.3 Feature selection

One of the primary advantages associated with the penalized regression modeling technique is that feature selection is an integrated aspect of model fitting. To determine the optimal level of regularization, the hyperparameters of both models were tuned using stratified 5-fold cross-validation. After identifying the optimal value for the tuning parameter, the final models were retrained on the full training sets using the chosen hyperparameters. The logistic model trained with the optimal C on the full training set inherently performed feature selection due to the LASSO penalty. Many SNP coefficients were shrunk all the way to zero, eliminating those features from the model. We examined the fitted model's coefficients and counted the number of SNP features with non-zero weights. The resulting non-zero fixed effects then corresponded to SNPs retained in the model. The remaining SNPs were shrunk to zero and discarded from the analysis through the regularization process. The substantial reduction in the number of features demonstrates that LASSO imposes sparsity and identifies a selective set of predictors. The number of selected SNPs for the lasso model with the optimal value C=3.16228 is shown in Table 5.3.

Table 5.3: The number of selected SNPs for the lasso model with optimal parameter.

Population	**Selected SNPs**	**C**
AA-F	1,271	3.16228
AA-M	1,209	3.16228
EA-F	1,514	3.16228
EA-M	1,890	3.16228

These values show that lasso reduced the size of the predictor space, and still managed to keep enough variants reflecting the polygenic nature of schizophrenia.

For group lasso, the selected feature counts with the optimal values of λ were 0.1 in all subgroup-specific thresholds also selected during tuning shown in Table 5.4.

Table 5.4: The number of selected SNPs for the group lasso model with optimal parameter.

Population	**Selected SNPs**	λ
AA-F	1,530	0.1
AA-M	1,064	0.1
EA-F	1,199	0.1
EA-M	1,040	0.1

Despite both approaches achieving substantial dimensionality reductions, the lasso regression model was selected as the best analytical framework for future interpretation due to its superior and consistent CV performance across each of the four subpopulations. That consistency is important. In high-dimensional genetic data, a method that performs consistently across matched analyses is usually more reliable than one that has an excellent performance in a single subgroup but has a vastly different performance in another subgroup.

Chapter 6

Results and Discussion

This chapter represents more detailed findings from the GWAS analyses of schizophrenia that were performed within this study based on two separate datasets: the GAIN and MGS_nonGAIN datasets. The methods described in Chapter 3 and Chapter 4 were utilized in both datasets, with the findings from the analyses of GWAS and the results of external validation and replication with MGS_nonGAIN will be presented. Finally, the findings from the LASSO and Group LASSO algorithms, including their predictive accuracy and biological relevance, will also be provided.

6.1 Results of prediction of schizophrenia for GAIN data

6.1.1 The test-set performance of the population-specific models

The predictive performance presented in this chapter was obtained using the methodological framework described in Chapter Chapter 3, in which strict separation of training and test data, feature selection within stratified 5-fold cross-validation, and independent test evaluation were implemented to minimize overfitting and prevent information leakage. The AUC, accuracy, sensitivity, and specificity for each trained ethnicity- and gender-specific RF model on the test set is presented in Table 6.1 below. We have provided a 95% CI to help assess the robustness of the AUC estimates. We have added 95% CI for the AUC values using the non-parametric bootstrap method (1,000 resamples) on the independent test set. We believe this addition offers a clearer interpretation of the model's discriminative performance across subgroups.

Table 6.1: The AUC (95% CI using 1000 bootstrap resamples), accuracy, sensitivity and specificity of the RF model on the independent test set (adapted from [56]).

RF	AUC	Accuracy	Sensitivity	Specificity
AA-F	69.5% (95% CI: 63.4%–75.2%)	68.6%	85.6%	41.1%
AA-M	74.0% (95% CI: 68.1%–79.4%)	73.9%	39.7%	93.3%
EA-F	69.2% (95% CI: 63.3%–75.0%)	75.1%	95.4%	34.2%
EA-M	74.3% (95% CI: 63.3%–75.0%)	65.4%	73.6%	58.6%

The results demonstrated that ethnicity-gender-specific trained random forests perform extremely well against on corresponding test sets. This finding is the highest level of performance achieved to date in predicting schizophrenia using SNP profiles. Table 6.2 provides the confusion matrices produced by the constructed RF model.

Table 6.2: Confusion matrices for each ethnicity and gender group, trained RF classifier evaluated on their own test sets (adapted from [56]).

		True class: case	True class: ctrl	
Predicted class	Positive	101 (TP)	43 (FP)	PPV 70.1%
	Negative	17 (FN)	30 (TN)	NPV 63.8%
		Sensitivity 85.6%	Specificity 41.1%	Accuracy 68.6%
		191		

AA Female

		True class: case	True class: ctrl	
Predicted class	Positive	27 (TP)	8 (FP)	PPV 77.1%
	Negative	41 (FN)	112 (TN)	NPV 73.2%
		Sensitivity 39.7%	Specificity 93.3%	Accuracy 73.9%
		188		

AA Male

		True class: case	True class: ctrl	
Predicted class	Positive	146 (TP)	50 (FP)	PPV 74.5%
	Negative	7 (FN)	26 (TN)	NPV 78.8%
		Sensitivity 95.4%	Specificity 34.2%	Accuracy 75.1%
		229		

EA Female

		True class: case	True class: ctrl	
Predicted class	Positive	103 (TP)	70 (FP)	PPV 59.5%
	Negative	37 (FN)	99 (TN)	NPV 72.8%
		Sensitivity 73.6%	Specificity 58.6%	Accuracy 65.4%
		309		

EA Male

6.1.2 Mapping single-nucleotide polymorphisms to genes

Then we mapped the identified variants (SNPs) to corresponding gene loci using public resources such as GeneCards [126] and the UCSC Genome Browser database [127]

to confirm biological relevance. Mapping SNPs to genes is a common step in genetic research - especially for genome-wide association studies (GWAS), variant annotation, and functional genomics analyses. It generally involves identifying which gene(s) are potentially affected by a given SNP. Furthermore, the same SNPs were also mapped to the corresponding genes by using the Gene Data Viewer tool [111] from the NCBI provided by the Affymetrix SNP Array 6.0 platform. This process yielded a set of genes for each group that significantly contributed to predicting the phenotype. Several selected genes, such as ARHGAP26, TEK, SLC44A1, and ERCC6 have been previously associated with schizophrenia-related biological processes, including neurodevelopment, synaptic signaling, and immune function, all of which are biologically relevant to schizophrenia [24, 42, 94, 128–134]. However, we also recognized 10 genes (MTCO1P53, TUBB4BP8, BAG1P1, LOC107986270, LOC10192694, LOC107985674, RPL21P71, RNA5SP181, C12orf45, and ATG4AP1) that have not been reported before. The selected 12 SNPs identified by the CMH association test for each ethnicity-gender population, their DNA locations, and corresponding genes are presented in Table 6.3. In our findings, these SNPs were identified in at least two subpopulations (see Table 6.3). Also, other identified SNPs with the mapped genes for each ethnicity-gender population are provided in Appendix Table A.5.

Table 6.3: Common 12 SNPs and genes among four ethnicity-gender populations (from [56]).

#	Chr	Position	SNP ID	Between ethnicity-gender	Gene
1	3	chr3:8872402	rs11131152	EA-F and EA-M	RAD18
2	5	chr5:143170695	rs224494	EA-F and EA-M	ARHGAP26
3	6	chr6:45040800	rs588291	EA-F and EA-M	SUPT3H
4	8	chr8:145041048	rs7816088	EA-F and EA-M	PLEC
5	9	chr9:27162125	rs10116212	EA-F and EA-M	TEK
6	9	chr9:107117818	rs10991619	EA-F and EA-M	SLC44A1
7	10	chr10:27707473	rs2992009	EA-F and EA-M	MKX
8	10	chr10:50351163	rs4253197	AA-F and AA-M	ERCC6
9	11	chr11:78741369	rs498820	EA-F and EA-M	TENM4
10	20	chr20:3535400	rs658709	EA-F and EA-M	ATRN
11	22	chr22:45486452	rs2111399	EA-F and EA-M	CERK
12	23	chr23:89022317	rs2552707	AA-M and EA-M	TGIF2LX

We also examined the functional categories of the SNPs and found that intronic SNPs were the most predictive, consistent with what is already known about their role in gene regulation. Our methodology highlights the importance of keeping feature selection, model training, and validation as separate steps. This separation is a key part of our study design and helps ensure that the models are reliable, perform well on new data, and are not overfit.

6.1.3 Prognostic impact of different types of SNPs

We performed a predictive analysis to evaluate the potential causal relationships between different types of SNPs, such as those found in introns, exons, missense, upstream, and downstream regions. For each SNP-type classification, we created a random forest classifier model for each population-specific dataset (because RF was the best model in model selection in the Methodology section), where a specific SNP type was used to train the classifier. We used a 5-fold cross-validation procedure during hyperparameter optimization of each random forest model for each SNP type. The results from the analysis of the SNPs indicated that intronic SNPs are the most predictive of schizophrenia (see Table 6.4). This was not surprising given that the majority of the SNPs identified for each population were intron SNPs. The predictive accuracy measured by AUROC for 5-fold CV and test dataset. At the bottom, the accuracy of the models designed in different scenarios is presented in Table 6.4.

Table 6.4: Prediction performance of different types of SNPs with RF model for a) AA-F, b) AA-M, c) EA-F and d) EA-M individuals.

Type	SNP	CV Accuracy	CV AUC	Test Accuracy	Test AUC
Intron	**6**	**72.9%**	**71.8%**	**67.2%**	**69.5%**
Upstream	1	64.1%	61.0%	63.9%	57.6%

a) AA-F – Random Forest model (max depth: 20; n_estimator: 100)

Type	SNP	CV Accuracy	CV AUC	Test Accuracy	Test AUC
Intron	**4**	**69.8%**	**67.7%**	**74.4%**	**64.6%**
Upstream	3	68.2%	61.8%	68.1%	59.9%

b) AA-M – Random Forest model (max depth: 10; n_estimator: 10)

Type	SNP	CV Accuracy	CV AUC	Test Accuracy	Test AUC
Intron	**14**	**82.4%**	**83.7%**	**71.2%**	**68.9%**
Upstream	2	72.7%	56.9%	67.3%	57.9%
Downstream	1	70.8%	54.0%	66.1%	53.9%
Missense	1	70.4%	55.0%	65.1%	53.9%

c) EA-F – Random Forest model (max depth: 50; n_estimator: 100)

Type	SNP	CV Accuracy	CV AUC	Test Accuracy	Test AUC
Intron	**21**	**71.1%**	**81.2%**	**65.4%**	**68.3%**
Upstream	5	62.4%	68.7%	57.6%	63.6%
Downstream	10	66.4%	73.9%	64.7%	71.6%
Exon	1	56.3%	55.0%	54.6%	55.5%

d) EA-M – Random Forest model (max depth: 50; n_estimator: 100)

6.1.4 Cross-population evaluation

Cross-population evaluation (CPE) is a type of evaluation to assess the effectiveness and accuracy of a given model within a specific population, based on performance data from another population. This method allows us to identify biases that arise when generalizing results to different populations. It provides researchers with valuable insight into how well-developed predictive models are when applied to other groups. CPE will be highly valuable for genomics research, as genetic and environmental factors between populations may produce significantly different results across various groups.

To assess the ability of trained RFs for the ethnicity-gender group explored, we performed a complementary analysis to determine the effectiveness of each RF's ethnicity-gender combination when applied to other populations. The results of the cross-population analysis are shown in Table 6.5 in terms of AUC and accuracy. This clearly demonstrates that almost all trained random forests show a significant decrease in predictive power when applied to other populations, thus confirming the specificity of each trained random forest to its own population.

Table 6.5: The AUC and accuracy of each ethnicity-gender-specific RF model when evaluated on test sets from other populations (adapted from [56]).

Population	**AA-F**		**AA-M**		**EA-F**		**EA-M**	
Model	**AUC (%)**	**Acc (%)**	**AUC (%)**	**Acc (%)**	**AUC (%)**	**Acc (%)**	**AUC (%)**	**Acc (%)**
AA-F	**69.5**	**68.6**	68.2	65.9	55.4	63.7	53.9	56.9
AA-M	70.0	72.6	**74.0**	**73.9**	60.5	68.6	68.9	71.2
EA-F	68.2	69.7	68.0	68.5	**69.2**	**75.1**	67.6	68.3
EA-M	70.8	64.6	69.2	63.6	65.4	62.7	**74.3**	**65.4**

6.2 Results of external validation

External validation analyses allow us to evaluate whether the predictive structure obtained in GAIN remains stable when applied to an independent but genetically matched cohort. So, we validated the results, assessed the model's performance and compared the model's performance to our original results. The results of external validation are presented in Table 6.6.

Table 6.6: The AUC (95% CI using 1000 bootstrap resamples), accuracy, sensitivity and specificity of each ethnicity-gender-specific RF model on the external non-GAIN data.

RF	AUC (95% CI)	Accuracy	Sensitivity	Specificity
AA-F	76.2% (59.3%–86.7%)	45.1%	41.3%	80.0%
AA-M	58.4% (50.8%–75.0%)	42.5%	31.4%	75.0%
EA-F	62.5% (63.3%–75.0%)	55.6%	85.6%	18.3%
EA-M	50.2% (47.1%–53.0%)	50.4%	55.2%	45.3%

While there was a decrease in accuracy and sensitivity for AA-females in the external sample, but AUC increased. One likely reason for this is that there was only a small sample size in the external sample (n = 51). This decrease may also have been partly due to the different size of the external sample, which lowers statistical power. The results from the performance of the male subgroups and EA female subgroups decreased from their respective GAIN evaluation to the external population because of cohort-specific differences in the distribution of allele frequencies (see Section 6.5 below).

6.3 Results of external replication for MGS_nonGAIN data

The results of the external replication analyses that we discussed in Chapter 4 and particularly in Section 4.2: *"External replication analyses"* include SNPs identified from the GAIN cohort in the AA-F and EA-F cases. For instance, 27 SNPs passed the FDR < 0.1 threshold in the AA-F case and 5 SNPs of them (more than 50%) matched the SNP set identified through the GAIN-trained model. This observation clearly supports the validity of the SNPs identified by our models. In the same way, the EA-F subgroup also had 36 SNPs pass the same cutoff level, of which 18 SNPs (100%) overlapped with the selected SNPs from GAIN data. However, no SNPs could be modeled for the AA-M and EA-M subgroups due to FDR < 0.1 (the minimum FDR is ≈ 0.9) not falling within statistical power. Table 6.7 shows the results of the RF external replication study for the test group from these studies. For the AA-M and EA-M subgroups, not a single SNP passed the FDR < 0.1 test (the minimum FDR is ≈ 0.9), preventing model construction. Here, we see that the confidence interval is incomplete due to the limited sample size.

Table 6.7: RF model performance on the test set for the external non-GAIN data (adapted from [56]).

RF	AUC	Accuracy	Sensitivity	Specificity
AA-F	71.4%	63.6%	50.0%	80.0%
EA-F	51.5%	60.2%	62.1%	47.6%

As we discussed before, to identify the discrepancy for AA-M and EA-M, we calculated allele frequencies and found no SNPs with FDR < 0.1. One example of how two datasets can yield very different observations through comparisons of minor allele frequencies is the comparison between the GAIN and MGS_nonGAIN datasets. In the GAIN dataset, we can see minor alleles that are more common in cases than controls, and in the MGS_nonGAIN dataset, the same minor alleles are more common in controls than cases. These differences can occur based on differences in sample sizes, population structure, or random chance within a smaller sample. The GAIN dataset showed significant associations between 37 SNPs (EA-M subgroup) at an FDR of less than 0.1, while there were no associations observed in the MGS_nonGAIN dataset, which emphasizes that the associations are very sample-dependent.

The significance of these findings has been attributed to the presence of randomly overrepresented SNP variants appearing more frequently in cases in one of the datasets solely due to sampling composition, rather than actual biological association with the disease. In our current analysis, we evaluated the same SNPs present across both datasets, evaluated the case-control frequency differences of those SNPs between the two study populations, and determined that in GAIN, there were some SNPs strongly enriched in cases, whereas in MGS_nonGAIN the SNPs were not, and vice versa.

The patients included within the datasets used do not include any overlap by population subtype. There was observed no overlap within the SNPs only when looking at the AA-female vs. EA-female. Contrarily, comparisons made between AA and EA males show no SNPs overlapping at a FDR < 0.1, with all SNPs present in the data of the nonGAIN population. Therefore, the allele frequency profiles between these two populations (AA-M and EA-M) are substantially different from one another.

6.4 Results of lasso and group lasso regression analysis

The implementation and results of a schizophrenia classification pipeline using LASSO and Group LASSO-regularized logistic regression models on GAIN GWAS data were provided. The final model with the selected SNPs for each subgroup and their learned coefficients was evaluated on the held-out test set. Accuracy, AUC, sensitivity, and specificity were used to assess model performance on the test set. Table 6.8 and

Table 6.9 show the evaluation performance of the lasso and group lasso models on the test set.

Table 6.8: Lasso model performance on the test set for the GAIN data.

Population	AUC	Accuracy	Sensitivity	Specificity
AA-F	83.8%	75.3%	61.1%	83.6%
AA-M	83.5%	75.6%	83.2%	63.5%
EA-F	83.2%	82.0%	64.4%	90.3%
EA-M	85.4%	77.0%	77.6%	76.1%

Table 6.9: Group lasso model performance on the test set for the GAIN data.

Population	AUC	Accuracy	Sensitivity	Specificity
AA-F	91.5%	86.6%	79.2%	91.0%
AA-M	85.6%	78.2%	94.1%	52.7%
EA-F	77.3%	68.3%	86.3%	50.3%
EA-M	78.1%	65.5%	87.6%	43.3%

The AUC for the lasso model ranged from 83.2% to 85.4% while the overall accuracy range was between 75.3% - 82.0%. The maximum AUC was found in the EA-M (85.4%) subgroup and the greatest accuracy was associated with the EA-F (82.0%) subgroup. The sensitivity and specificity also showed an evenly distributed classification performance, with some subgroup-specific variation present. Group lasso showed mixed results on the test set. In AA-F, group lasso produced strong performance, with an AUC of 91.5% and an accuracy of 86.6%. However, in the other subgroups, its performance was less stable, particularly with respect to specificity, which dropped to 52.7% in AA-M, 50.3% in EA-F, and 43.3% in EA-M. Overall, although group lasso performed well in one subgroup, the lasso approach demonstrated more consistent behavior across all four populations.

The selected SNPs in the LASSO model were then mapped to nearby genes for the purpose of examining biological significance. The resulting gene counts and overlaps are summarized in Table 6.10 .

Table 6.10: Lasso-based SNP selection: subpopulation-specific interpretation and gene overlap analysis.

Comparison	**Gene Count**
AA Female genes	711
AA Male genes	663
EA Female genes	835
EA Male genes	1015
AA Female ∩ AA Male	83
AA Female ∩ EA Female	108
AA Female ∩ EA Male	131
AA Male ∩ EA Female	99
AA Male ∩ EA Male	120
EA Female ∩ EA Male	178
Overlap across all 4 groups	**16**

At the gene level, there were both specific signals of the subgroups and their shared genetic uniqueness from one another. While there were numerous unique genes per subgroup. There was also an overlap of genes across populations, with a total of 16 genes overlapping all four of the subgroups. These results support the conclusion that there is no single cause for schizophrenia; rather, it is caused by a combination of both common polymorphic and population-specific genetic variants. The existence of 16 genes common to all groups' points to a core set of loci that may contribute broadly to schizophrenia risk, while the larger number of subgroup-specific genes indicates heterogeneity related to ancestry and sex. That is not surprising in a disorder as genetically complex as schizophrenia, but it is still useful to see it emerge so clearly from the selected features.

We performed pairwise linkage disequilibrium (LD) analysis further to characterize the genomic architecture of the lasso-selected SNPs using the PLINK tool. The purpose of this analysis was to evaluate correlations among the predictive SNPs within each ethnicity and gender subgroup. The R^2 metric was used to measure LD, with higher values indicating greater co-inheritance among SNPs.

The LD analysis for the different subpopulations presents in Table 6.11.

Table 6.11: LD analysis for all subpopulations.

Population	**Total Pairs**	**Mean** R^2	**Median** R^2	**Max** R^2	$R^2 > 0.8$	$R^2 > 0.5$
AA-F	3,770	0.279	0.233	1.000	27	221
AA-M	1,956	0.249	0.206	0.951	7	58
EA-F	2,911	0.330	0.273	0.948	36	388
EA-M	1,956	0.249	0.206	0.951	7	58

As shown in Table 6.11, the EA-Females case had the highest average LD value (mean R^2 = 0.33) and the greatest number of strong correlations between SNP pairs

(388 pairs with $R^2 > 0.5$), indicating a greater concentration of haplotype blocks. Both AA-M and EA-M had lower overall LD values, indicating lower correlation between nearby SNPs. Differences in LD structure across different populations can influence both the sparsity of the model and its interpretation.

The findings presented throughout this chapter have illustrated that sparsely regularized logistic regression analysis is an effective analysis for predicting an individual's diagnosis of schizophrenia using high-dimensional GWAS data. The Lasso provides a strong and consistent classification framework for schizophrenia prediction using high-dimensional GWAS data. These findings are comparative when compared with previously published results from research on schizophrenia utilizing machine learning approach. For example, multimodal machine learning models that incorporate genetic and imaging data report AUC values from 0.76 to 0.92 [135], whereas there are other classification methodologies that utilize clinical and imaging features that have reported an approximately 90% accuracy but have been developed using different data and feature sets [136]. The research conducted by Ma et al. [137] established that LASSO was superior to standard logistic regression and various other classification methods for disease prediction through genetic association studies (GWAS) when a large number of SNPs, which were highly correlated, were present. By performing simultaneous shrinkage and feature selection, LASSO reduced the effect of noise from non-informative genetic variants; thus, allowing for an increase in the accuracy of the classifications. Work conducted by Wu et al. [138] has been presented to indicate that feature selection and enhancing predictive power via GWAS data through LASSO penalised logistic regression provides improved results over traditional methods. This is supported in the current study through demonstration of a greater AUC value for LASSO compared to other algorithms, thereby illustrating the advantage obtained through combining techniques of regularisation with tailored preprocessing and subgroup analysis. Another major finding of this study is associated with the ability of biological interpretability of the models. Unlike ML methods based on feature extraction or transformation, the use of embedded feature selection methods allowed the original SNP variables to be preserved. This allowed us to directly map the selected variants to genes and execute overlap analysis across different subpopulations. The identification of both shared genes and subgroup-specific genes which can contribute to elucidate the polygenic and heterogeneous nature of schizophrenia.

6.5 Discussion

This research was conducted to establish the capability of high-dimensional machine-learning approaches to predict risk for schizophrenia disorder using SNP data across the entire genome. Moreover, it also measured the strength and applicability of predictive

models for schizophrenia across independent data sources. The results provide many valuable insights regarding both methodological issues associated with predicting schizophrenia and biological influences on the prediction of schizophrenia.

The choice of Random Forest in this study was determined by the characteristics of the available GWAS SNP-array data, the study objectives, and the need to preserve the biological interpretability of SNP predictors rather than by the complexity or novelty of the classification algorithm. Considering the relatively limited sample sizes after ethnicity- and gender-specific stratification, together with the high dimensionality of the SNP data, random forest provided a robust and appropriate solution for the objectives of this research. The predictive performance observed across the analyzed datasets further supports the suitability of Random Forest for structured high-dimensional GWAS data. These findings demonstrate that classical machine learning approaches remain effective for SNP-array-based schizophrenia prediction under the characteristics of the datasets analyzed in this study.

Numerous genetic factors commonly underpin complex traits such as schizophrenia, each having a small effect size [151]. During our investigations, it was shown that it is possible to predict schizophrenia using machine learning techniques based on the single-nucleotide polymorphisms (SNPs). Due to this, we stratified data according to ethnicity and gender to prevent demographic stratification [114–117]. As shown in the GAIN dataset analysis, in Chapter 3, after completing extensive model selection, a singular model was selected and constructed using training data of ethnicity-gender-based determinants. The results from the various trials of the model selection confirmed the random forest algorithm, in each trial, as the appropriate approach to create the selected model. The overall classification accuracy (AUC, sensitivity, and specificity) from the developed random forest models for four groups, AA-F, AA-M, EA-F, and EA-M, was 68.6% (69.5%, 85.6%, 41.1%), 73.9% (74.0%, 39.7%, 93.3%), 75.1% (69.2%, 95.4%, 34.2%), and 65.4% (74.3%, 73.6%, 58.6%), respectively. The classification accuracy was evaluated on independent test sets comprised of individuals of similar ethnicity and gender as those in the study. To our knowledge, these findings represent the highest prediction accuracy to date for schizophrenia using SNPs. Meaningful discrimination power because the genetic risk of developing schizophrenia is made up of many different risk variants with small individual effect sizes. Regarding the limited sample size, that said, schizophrenia is not a very common disease in the general population, which makes it harder to collect large and balanced datasets of cases and controls, especially for stratified analyses across ethnic ethnicity-gender subgroups. We know that the sample size in our study is not large, even though it is comparable to some other GWAS studies [94, 95].

To mitigate the challenges posed by sample size, we implemented 1) population-stratified modeling (by ethnicity-gender) to avoid confounding and improve signal-

to-noise ratio; 2) 5-fold cross-validation and held-out test set evaluation to reduce overfitting; and 3) FDR-controlled feature (SNP) selection to limit inclusion of noise variables. We used the BH-FDR method to control for the large number of tests and lower the risk of false positives when picking SNPs. This approach helped us strike a balance by identifying SNPs that are likely associated with schizophrenia while minimizing the inclusion of too many false positives.

We do not view our models as the ultimate predictors of the disease but rather as a proof of concept showing that SNP-based classification models can still achieve reasonable prognostic efficacy (AUC $\sim$ 69–74% and accuracy $\sim$ 65–75%), even though schizophrenia is highly polygenic, and our sample size is limited. These models still need to be tested on larger and more diverse datasets before they can be used in real clinical settings. The low specificity of the EA model means that sensitivity and specificity are less useful for accurate diagnosis. However, it may be useful for profiling an individual at an early stage, i.e., as a first step to identifying individuals who may require further examination.

We agree that schizophrenia involves many genetic variants (SNPs), each having a small effect size. Because of this, large sample sizes are usually needed to detect strong genome-wide associations. Our study included about 4,700 individuals, so we recognize that the power to identify new common SNPs meeting strict GWAS thresholds is limited.

While these SNPs may not meet classical GWAS thresholds, they could still be biologically meaningful and worth further study. In our study, we separate the data by ethnicity and gender, which helps us find genetic variants that might be missed in larger studies that mostly include European populations. It is also important to note that our goal is not to discover new associations, but to build predictive models using a combination of SNPs.

Nevertheless, we mapped the selected SNPs to genes and found several that had not previously been associated with schizophrenia in the GWAS literature. These may reflect population-specific variants, particularly within the African ancestry subgroups, which are historically underrepresented in psychiatric genetics studies. We acknowledge that these novels identified variants require replication and functional validation. We include this as a limitation and call for expanded studies in diverse populations to further explore ancestry-specific genomic risk factors. Furthermore, in our study, we focused on autosomal SNPs. We did not include sex chromosomes due to the complexities they introduce in GWAS, such as dosage differences, inheritance patterns, and platform issues, which make it more complicated.

In total, we identified 67 SNPs which used four trained classification models across four ethnic and gender groups. The 67 SNPs were then mapped to genes, which resulted in 12 common SNPs (rs11131152, rs224494, rs588291, rs7816088,

rs10116212, rs10991619, rs2992009, rs4253197, rs498820, rs658709, rs2111399) and common genes (RAD18, ARHGAP26, SUPT3H, PLEC, TEK, SLC44A1, MKX, ERCC6, TENM4, ATRN, CERK and TGIF2LX) that were associated with at least two out of the four ethnicity and gender-specific individuals. Due to the data-driven nature of our study design, we did not prioritize human leukocyte antigen (HLA) variants because they did not meet the BH-FDR threshold during our feature selection methodology. There are a few reasons why we didn't find HLA variants in our results. The main one is that there were not enough people in this study for researchers to find links between HLA variants and other things, especially ones that have smaller links or have specific effects on the population. Second, the multiple-testing correction that is rigorous (BH-FDR threshold of T<0.1) may require excluding genetic variants of HLA with associations weakened in the analytic model due to sample size or other variables unique to that sample. Lastly, due to our modeling strategy emphasizing ethnic and gender-specific subgroups (AA-F, AA-M, EA-F, EA-M), it is possible that we prioritized SNPs that had a stronger subgroup-specific association compared to SNPs with broader associations such as those observed within the HLA region.

In our models, several genes are previously identified as associated with schizophrenia. Examples of genes we found include RAD18, which is at 3p25.3 (distance=20978), ARHGAP26 at 5q31.3, SUPT3H at 6p21.1, TEK (also known as TIE2) at 9p21.2, and SLC44A1 at 9q31.1 (distance=305711), MKX on 10p12.1, ERCC6 at 10q11.23, TENM4 at 11q14.1, ATRN on 20p13 and CERK at 22q13.31, which have all been previously identified associated with a risk for developing schizophrenia [94]. Another study [128] identified a correlation between ARHGAP26 and schizophrenia based upon the results from investigations using DNA methylation [128]. The study also identified 2 genes involved in psychomotor behavior associated with schizophrenia can be found within the following 2 genes: PLEC and ERCC6 [129]. Previous studies have shown associations between TEK [130], SLC44A1 [130], ERCC6 [140], TENM4 [131], ATRN [141], CERK [132], and TGIF2LX [133, 134] genes and schizophrenia, which was also confirmed in our analysis. We have discovered 10 gene loci that have never been associated with schizophrenia in other studies. Therefore, this finding implies that the ethnicity-gender-specific models developed in this study are novel and currently relevant. Supporting evidence for this conclusion includes gene loci, including: MTCO1P53, TUBB4BP8, BAG1P1, LOC107986270, LOC107985674, RPL21P71, RNA5SP181, C12orf45, and ATG4AP1.

SNP rs4253197 has been noted as a covariance between BP and alcohol use through a cross-sectional design study [142, 143]. For schizophrenia and alcohol dependence [144], some genetic variants and loci may have evidence of covariance. This SNP could potentially be the first association of a novel SNV with BP traits [143] for individuals with African ancestry. Researchers have identified marker SNP rs9629977

using clustered SNPs found in GWA studies related to alcohol dependence conducted on subjects of African American ethnicity [145].

To analyze marker SNPs together, we first examined the relative weights functional categories of the SNPs and found that intronic SNPs were the most predictive, which matches what is already known about their role in gene regulation (see Table 6.4). Our methodology highlights the importance of keeping feature selection, model training, and validation as separate steps. This separation is an important part of our study design, and it helps ensure that the models are reliable, work well on new data and are not overfitted. After each model for ethnicity-gender-SNP was trained on the appropriate training sets, then the model was tested on their respective test sets. Results of the evaluation can be found in Table 6.12. The majority of identified SNPs in the results of Table 6.12 were found to be intron. Furthermore, they also produced the greatest amount of predictive capability for schizophrenia. The finding confirms previous research has indicated that intron regions play a vital role in predicting schizophrenia [23, 139, 146–148].

Table 6.12: The AUC values for functional categories of the SNP type model on the test set [56].

Type	**AA-F AUC**	**AA-M AUC**	**EA-F AUC**	**EA-M AUC**
Intron	69.5%	64.6%	68.9%	68.3%
Upstream	57.6%	59.9%	57.9%	63.6%
Downstream	–	–	53.9%	71.6%
Missense	–	–	53.9%	–
Exon	–	–	–	55.5%

For the purpose of avoiding batch effects and familial substructures, we also assessed pairwise genetic distances (pairwise identity-by-state (IBS) distances among all individuals within each subgroup. In our analysis, we found no evidence of such confounding. Instead, our data reveal significant genetic heterogeneity and an absence of familial clustering in the schizophrenia cohort. Table 6.13 shows the pairwise IBS distances.

Table 6.13: Summary of pairwise genetic distance [56].

Population	**Mean Distance**	**Standard Deviation**	**Minimum**	**Maximum**
AA Female	0.2483	0.0046	0.2179	0.2924
AA Male	0.2694	0.0039	0.2339	0.3151
EA Male	0.2571	0.0021	0.2380	0.2777
EA Female	0.2569	0.0019	0.2473	0.2777

The results show that the average pairwise genetic distance was high, with a small standard deviation. This means that a sample was genetically heterogeneous with no close relatives or tight sub-clusters. Histogram plotting pairwise distance in Fig. 6.1

indicates a single smooth distribution, and no pairwise distances approach zero, which would suggest close relatives or the presence of population substructure. Variation of AA cohorts (Fig. 6.1 a&b) appear to have somewhat greater variation than expected; this is consistent with the high amount of genetic variation found in populations of African ancestry. Conversely, there is limited variation in the subgroups of European ancestry (Fig. 6.1 c&d), as predicted, and thus, there is adequate variation within these subgroups and no evidence of family clustering or unusual genetics.

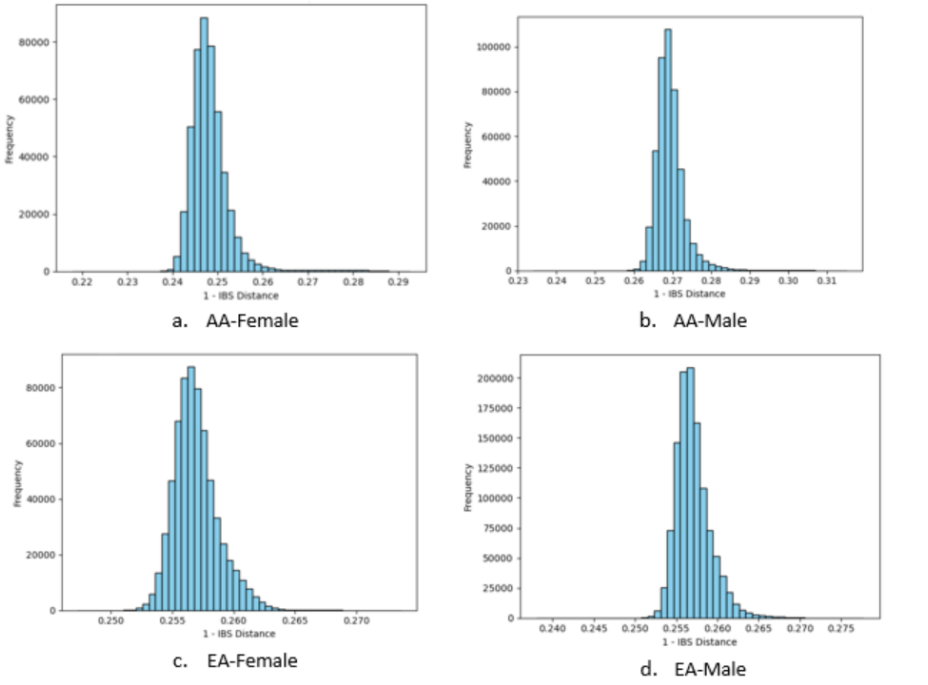

Figure 6.1: IBS pairwise genetic distances for each sub-population [56].

Moreover, in accordance with the GAIN data documentation, the phenotypic data are gathered from many clinical sites utilizing different assessment protocols. The presence of multiple methodologies significantly reduces the likelihood that the identified genetic signals are artifacts of or a single run of experiments or batch.

Thus, we can conclude that collectively, these results confirm our samples are genetically heterogeneous, do not contain large clusters of genetic relatedness, and that diagnoses were obtained in different clinics using various testing protocols. Therefore, the batch effect will not play any role.

Next, we used another schizophrenia SNP-based GWAS dataset to replicate our results from GAIN data which is presented in Chapter 4. The independent external dataset was Molecular Genetics of Schizophrenia - nonGAIN (MGS_nonGAIN) sample. The results of the external validation and replicated analysis give additional understanding to their interpretations. When using models derived from the GAIN database to test independently collected data (MGS_nongain), performance decreased compared to the corresponding model. It is known that genetic models do not

always transfer accurately between cohorts. Variations in allele frequency, linkage disequilibrium structure, diagnostic criteria, and the sample composition are just some examples of how predictive power can differ among datasets. The lower performance recorded in external validation shows the dependence of machine learning models on features of the particular data set being used and that there are limits to relying on only internal validation criteria. Indeed, the fact that some sub-cohorts have no overlapping statistically significant SNPs within a specified FDR cut-off indicates the degree to which these signals are dependent upon the specific dataset being studied. The presence of a significant degree of overlap between the AA-female and EA-female subgroups provides some degree of reassurance. It indicates that at a minimum, some of the identified variants are stable genetic signals and not merely random variations. However, the overall interpretation remains limited in terms of its ability to generalize across populations; this is a fundamental barrier to translating GWAS-derived predictive models into clinical settings.

In support of the validity of the original results, we performed a replication analysis with the resampling method of the MGS_nonGAIN (nonGAIN) training set. While the predictive performance of the classifiers was not as strong as that seen with GAIN (i.e., the predictive accuracy of the benchmark classifier was lower than it had been during the initial analyses), random forest was once again determined to be the best performing classifier. This consistency indicates that the algorithm itself is robust to different datasets, even when the specific predictive features differ.

The external validation and the replication analyses highlight both the strengths and weaknesses of this proposed method. On one hand, they show that machine learning algorithms can capture important genetic information associated with schizophrenia.

The reduction in predictive performance observed during external validation should be interpreted in the context of evaluating the developed models on an independent cohort rather than as evidence of model overfitting. Although both the GAIN and MGS datasets investigate schizophrenia, they differ in participant recruitment strategies, diagnostic criteria, ancestry composition, allele frequency distributions, and sample characteristics. Furthermore, schizophrenia is a highly polygenic and genetically heterogeneous disorder, in which numerous variants contribute only small individual effects. Consequently, predictive SNP sets identified in one cohort are not expected to transfer perfectly to another independent population. Therefore, a moderate decrease in predictive performance during external validation is an anticipated outcome that reflects biological and cohort-specific variability rather than limitations of the proposed modeling framework. Importantly, despite the observed reduction in predictive performance, the consistent application of the proposed analytical pipeline across independent datasets demonstrates the robustness of the methodology. The external validation and independent replication analyses therefore provide evidence that the

proposed framework remains applicable across different cohorts, even though the specific predictive variants and their effect sizes may vary between populations.

The research findings support the common biological viewpoint that schizophrenia is a highly polygenic and heterogeneous disorder, which is well-known in the field. The locating of both previously identified genes, and possibly unique loci provides further support for the idea that there are multiple genetic pathways involved in the risk of developing the disorder.

While RF classifier showed the best performance in the main pipeline of GAIN data, the predictive performance of the lasso-based models during Chapter 5 showed that differed modeling types, capture different genetic features from DNA. This suggests that different modeling paradigms can reveal different aspects of the genetic signal. Tree-based techniques are the better choice for modeling nonlinear interactions among SNPs, while LASSO would allow one to find multiple sparsely populated interpretable solution sets by selecting a smaller list of the predictive variant genes. For high-dimensional genome data, it seems that individual SNP level sparsity through lasso does better than a group lasso penalty system. Group lasso model, while interesting from a biological standpoint, did not perform consistently across different subpopulations. This difference might be due to the fact that schizophrenia associated variants are dispersed throughout the genome rather than nicely grouped into some predetermined category. Therefore, biology does not always conform to the organization structure we impose in our models.

Compared with previous schizophrenia studies that incorporated lasso, our lasso test performance is clearly competitive and, in several cases, stronger. Chen et al. [99] used a lasso model to select polygenic score (PGS) predictors and then built downstream prediction models. In their combined MGS+SSCS sample, the best model achieved an accuracy = 0.813 and an AUC = 0.905, while in external CATIE validation, it dropped to an accuracy = 0.721 and an AUC = 0.747. Relative to those results, our lasso models show similar or slightly higher accuracy on the held-out test sets in some subgroups and consistently higher test AUC than their external validation results. Their best internal AUC remains somewhat higher than our held-out GWAS SNP-based lasso test AUCs. This is actually a meaningful comparison because their study did not use raw SNP predictors in the same way, but rather lasso-selected PGS-based predictors.

A second important comparison is the study by Wang et al. [149], which created logistic and lasso-based nomograms to estimate the risk of schizophrenia using minor physical abnormalities (MPAs) instead of genomic variants. Their lasso validation performance was AUC = 0.85 and accuracy = 80.5%, with sex-stratified AUCs of 0.89 in males and 0.91 in females. Against that benchmark, our lasso model is remarkably close in overall test accuracy and very similar in AUC, despite relying on a much harder input space, namely high-dimensional GWAS SNP data instead of phenotypic anomaly measures. In that sense, our results support the idea that sparse genomic models can

reach performance levels comparable to non-genetic schizophrenia classifiers while retaining direct biological interpretability at the SNP level.

For group lasso model, direct schizophrenia diagnosis studies with fully reported AUC and accuracy are even rarer. One of the clearest related examples is the sparse-group-lasso neuroimaging study by Xiang et al. [150], where grouped feature selection was combined with an SVM classifier and achieved 93.1% accuracy for schizophrenia identification using fMRI-derived graph measures. Our best group lasso result, in AA-F (AUC = 0.915, accuracy = 86.6%), is somewhat lower than that neuroimaging-based accuracy benchmark, but it remains strong given that our model operates on GWAS data alone. At the same time, the fact that our group lasso performance drops substantially in the other subgroups suggests that group-structured sparsity is less stable than SNP-wise sparsity for this dataset, likely because schizophrenia risk variants are widely distributed and do not always align cleanly with predefined groups.

The comparison suggests three conclusions. First, our lasso results are strong by schizophrenia-prediction standards, especially because they are based on GWAS SNP data and evaluated on held-out test sets. Second, our group lasso results show promise but are less consistent, which agrees with the broader intuition that grouped penalties help only when the grouping structure reflects the underlying biology. Third, while some published studies report higher performance. Those models often use richer or less noisy modalities such as neuroimaging or clinical phenotypes. So, their results are not directly equivalent to SNP-only prediction. From that perspective, our lasso findings are not only competitive but methodologically important, because they show that interpretable sparse genomic models can achieve meaningful schizophrenia classification without abandoning the original biological features.

Chapter 7

Conclusions

In this concluding chapter of the thesis, we summarize the findings about the research questions set forth at the beginning (in Chapter 1), including future research opportunities.

7.1 Addressing the research questions

The primary aim of the thesis was to determine whether high-dimensional machine learning methods can effectively predict schizophrenia risk using GWAS SNP data. Additionally, determine whether these predictive models generalize across independent datasets. This study also examined the comparative performance of different type of modeling approaches (probabilistic models, ensemble methods, and regularized regression techniques) for predicting schizophrenia risk within this context. The obtained results provide answers to the research questions.

First, the results show that the schizophrenia risk can be predicted from GWAS SNP data with reasonable accuracy. Across the entire GAIN dataset, machine learning models demonstrated moderate but consistent predictive performance, confirming that genetic variation itself contains detectable signals associated with disease status. Among the evaluated methods, the RF model showed the best overall results, suggesting an important role of nonlinear relationships and interactions between SNPs in determining the schizophrenia risk.

The second part of this research indicates how algorithm selection and feature selection will highly affect how well a model performs. Regularized regression techniques also work well in high-dimensional feature spaces, especially lasso regularization. The models provided strong classification efficiency as well as a selecting list of informative SNPs. On the other hand, group lasso exhibited a more variable and less stable performance across the subpopulations, suggesting that group-based sparsity is not fully align of the distributed genetic architecture of schizophrenia.

The analysis further emphasizes the importance of feature selection in GWAS-based prediction. By preserving the original SNP variables, the proposed approach provides biological interpretability and allows for further analysis such as gene mapping and overlap analysis. Since schizophrenia is a complex disease, understanding the biological basis of prediction is just as important as achieving strong.

Fourthly, the analyses of the external validity and replication provide evidence on generalizability of models. On the independent MGS_nonGAIN dataset, the predictive performance of the models decreased, thus suggesting limited transferability across datasets. However, the consistent use of the RF model as the best classifier among both datasets indicates that certain modeling strategies remain robust even when the underlying genetic signals change. The results indicated that although predictive strategies can be generalized, the specific SNPs driving the predictions are often dataset dependent.

Lastly, the analysis of SNP functional categories revealed that intronic variants have the highest contribution to the prediction of schizophrenia.

In this thesis, we do not employ deep leaning techniques in schizophrenia prediction using SNP-based data. It was motivated by next considerations. First, one of the goals of this study was to maintain biological interpretability of the predictors. Since SNPs are specific types of genetic variants, when analyzing SNP data, it is important to keep their physical nature of our features, which are SNPs. For this reason, feature selection methods were preferred over feature extraction methods. In contrast with deep learning, where the model depends on automatic extraction and manipulation of features prior to transforming into a latent representation for prediction, the approaches analyzed here use original SNP variables, resulting in a much more interpretable output, and enabling downstream biological studies such as gene mapping. At the same time, and perhaps more importantly, our samples consist of genetic profiles of individuals. As a result, we do not have a sufficiently large sample size, which is generally required to train accurate deep learning models. Nevertheless, this does not mean that deep learning methods cannot be applied. It would still be possible to use dimensionality reduction techniques in combination with deep learning. However, the limited sample size remains a significant constraint.

To summarize, the identification of genetic variants is critical to enhancing the diagnosis of schizophrenia. Genomic data included in ML algorithms enables deeper understanding of the complex disorders like schizophrenia, thereby allowing significant advances to be made within the precision medicine domain. The ML-based predictive models enhanced in this thesis have the potential to be used as adjunctive clinical tools to assess the schizophrenia risk in individuals of EA and AA population using their SNPs.

7.2 Limitations

The thesis has several limitations. One such limitation is that the study was unable to control for any of the key risk factors outside of genetic risk factors, such as environmental factors (e.g., factors in early life that may be impacted by one's

environment), behavioral factors (e.g., the amount of smoking one does), and also, epigenetic factors (e.g., how the environment impacts the genetics of an individual). Therefore, the current models that we developed are not representative of all of the potential risk factors that are associated with the disease. Thus, there may be some limitations in terms of how effective the current models may be when applied in clinical practice. We used one of the largest publicly available the GAIN study from dbGaP databases on schizophrenia. However, the number of participants is still less than ideal for a genetic study of this nature. This limitation becomes more pronounced when dividing data by ethnicity and gender. Further reduces the sample size for each group, meaning the results may be less stable and reliable.

As a result of both the stratification of the sample and the polygenic nature of schizophrenia, it is possible that our results are subject to false positives especially with regard to searching for novel SNPs with small effect sizes. However, the BH-FDR method was implemented to control for multiple testing and reduce the likelihood of identifying false-positive SNPs; yet we acknowledge that many of the SNPs we have identified have not yet been validated biochemically. For instance, locations that have previously been associated with schizophrenia, like the HLA locus, were not seen in the results of our assumptions. This might suggest that our choice of texts creates a vulnerability to sample based signals.

This research illustrates that GWAS data-based predictive modeling techniques can still support complex psychiatric disorders, even with certain limitations. By merging genetic data with demographic covariate variables (e.g. gender, ethnicity), we were able to construct individual models for diverse subgroups. This may help better understand how schizophrenia manifests in different ethnicities. Predictive modeling over traditional GWAS approaches provides several advantages: 1) the identification of polygenic signatures involving both nonlinear and higher order interactions between markers, and 2) suitability to translational applications where a prediction is desired for risk at the individual level. Also, population stratification, stringent feature selection, cross-validation, and separate test evaluation are key components of our methods which provide a valid pathway to evaluate predictive power without giving any opportunity for data leakage or overfitting to occur.

7.3 Future studies

Future research needs to look at various types of biological information that can be obtained from multi-omics data like gene expression, epigenetic modifications, and environmental action. When combined together with GWAS SNP data, it will give a clear picture of what goes on with certain molecules that are involved in the pathogenesis of schizophrenia. These molecular data sets will allow for improving

predictive accuracy and potentially usefulness for patient care. In future research, we could also develop a unified view of genotype and phenotype through a combination of data mining methodologies for optimal usage within a single decision support system. Also, future directions might be developing an integration of a unified view of genotype and phenotype through a combination of data mining methodologies for optimal usage within a single decision support system.

Future studies can build upon the proposed framework by employing multi-omics data and multimodal deep learning methods together with genome-wide SNP information. The application of multimodal deep learning together with genomic foundation models and the integration of multiple omics disciplines may facilitate the discovery of latent genetic mechanisms underlying schizophrenia and improve predictive performance when sufficiently large multi-omics datasets become available.

Another important area is to develop models in which population heterogeneity is explicitly considered. The use of transfer learning, domain adaptation, and multi-task learning are examples of techniques that may improve the generalizability of models from one set of data or population to another.

Finally, extending the analysis to clinical practice, such as early detection of disease risk and personalized treatment plans will be critical for the eventual application of our findings. This will involve not only enhancing the level of predictive accuracy but also carefully considering ethical, clinical, and population-specific considerations.

Bibliography

[1] V. Tam, N. Patel, M. Turcotte, and et al., "Benefits and limitations of genome-wide association studies," *Nature Reviews Genetics*, vol. 20, pp. 467–484, 2019.

[2] Cross Disorder Phenotype Group of the Psychiatric GWAS Consortium, "Dissecting the phenotype in genome-wide association studies of psychiatric illness: cross-disorder phenotype group of the psychiatric gwas consortium," *British Journal of Psychiatry*, vol. 195, no. 2, pp. 97–99, 2009.

[3] W. S. Bush and J. H. Moore, "Chapter 11: Genome-wide association studies," *PLoS Computational Biology*, vol. 8, no. 12, p. e1002822, 2012.

[4] N. Craddock and L. Mynors-Wallis, "Psychiatric diagnosis: impersonal, imperfect and important," *British Journal of Psychiatry*, vol. 204, no. 2, pp. 93–95, 2014.

[5] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang, "Five years of gwas discovery," *American Journal of Human Genetics*, vol. 90, no. 1, pp. 7–24, 2012.

[6] J. Gratten, "Rare variants are common in schizophrenia," *Nature Neuroscience*, vol. 19, pp. 1426–1428, 2016.

[7] M. J. Taylor, J. Martin, Y. Lu, and et al., "Association of genetic risk factors for psychiatric disorders and traits of these disorders in a swedish population twin sample," *JAMA Psychiatry*, vol. 76, no. 3, pp. 280–289, 2019.

[8] K. S. Kendler, "Reflections on the relationship between psychiatric genetics and psychiatric nosology," *American Journal of Psychiatry*, vol. 163, no. 7, pp. 1138–1146, 2006.

[9] J. Gelernter, H. Kranzler, R. Sherva, and et al., "Genome-wide association study of alcohol dependence: significant findings in african- and european-americans including novel risk loci," *Molecular Psychiatry*, vol. 19, pp. 41–49, 2014.

[10] D. Petric, "Explaining schizophrenia from medical and philosophical perspective," *Open Journal of Medical Psychology*, vol. 11, pp. 191–204, 2022.

[11] K. Benallel, W. Mansouri, J. Salim, and et al., "Prescribing habits of moroccan psychiatrists toward patients with schizophrenia: about 72 practitioners," *Middle East Current Psychiatry*, vol. 30, p. 41, 2023.

[12] A. Shanko, L. Abute, and T. Tamirat, "Attitudes towards schizophrenia and associated factors among community members in hossana town: a mixed method study," *BMC Psychiatry*, vol. 23, p. 80, 2023.

[13] World Health Organization, "Schizophrenia." Available: https://www.who.int/news-room/fact-sheets/detail/schizophrenia, 2023.

[14] M. J. Owen, A. Sawa, and P. B. Mortensen, "Schizophrenia," *Lancet*, vol. 388, no. 10039, pp. 86–97, 2016.

[15] B. Mieth, M. Kloft, J. Rodríguez, and et al., "Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies," *Scientific Reports*, vol. 6, p. 36671, 2016.

[16] Schizophrenia Working Group of the Psychiatric Genomics Consortium, "Biological insights from 108 schizophrenia-associated genetic loci," *Nature*, vol. 511, pp. 421–427, 2014.

[17] C. Crisafulli, A. Drago, M. Calabrò, E. Spina, and A. Serretti, "A molecular pathway analysis informs the genetic background at risk for schizophrenia," *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 59, pp. 21–30, 2015.

[18] H. Yu, H. Yan, J. Li, and et al., "Common variants on 2p16.1, 6p22.1 and 10q24.32 are associated with schizophrenia in han chinese population," *Molecular Psychiatry*, vol. 22, pp. 954–960, 2016.

[19] H. Stefansson, R. Ophoff, S. Steinberg, and et al., "Common variants conferring risk of schizophrenia," *Nature*, vol. 460, pp. 744–747, 2009.

[20] Y. Shi, Z. Li, Q. Xu, T. Wang, and et al., "Common variants on 8p12 and 1q24.2 confer risk of schizophrenia," *Nature Genetics*, vol. 43, pp. 1224–1227, 2011.

[21] The International Schizophrenia Consortium, "Common polygenic variation contributes to risk of schizophrenia and bipolar disorder," *Nature*, vol. 460, pp. 748–752, 2009.

[22] A. F. Pardiñas, P. Holmans, and et al., "Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection," *Nature Genetics*, vol. 50, pp. 381–389, 2018.

[23] M. O'Donovan, N. Craddock, N. Norton, and et al., "Identification of loci associated with schizophrenia by genome-wide association and follow-up," *Nature Genetics*, vol. 40, no. 9, pp. 1053–1055, 2008.

[24] Z. Li, J. Chen, H. Yu, L. He, and et al., "Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia," *Nature Genetics*, vol. 49, pp. 1576–1583, 2017.

[25] M. Lam, C. Chen, and et al., "Comparative genetic architectures of schizophrenia in east asian and european populations," *Nature Genetics*, vol. 51, pp. 1670–1678, 2019.

[26] M. Hamshere, J. Walters, R. Smith, and et al., "Genome-wide significant associations in schizophrenia to itih3/4, cacna1c and sdccag8, and extensive replication of associations reported by the schizophrenia pgc," *Molecular Psychiatry*, vol. 18, pp. 708–712, 2013.

[27] The International Schizophrenia Consortium, "Common polygenic variation contributes to risk of schizophrenia and bipolar disorder," *Nature*, vol. 460, pp. 748–752, 2009.

[28] The Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium, "Genome-wide association study identifies five new schizophrenia loci," *Nature Genetics*, vol. 43, pp. 969–976, 2011.

[29] P. J. Harrison, "Recent genetic findings in schizophrenia and their therapeutic relevance," *Journal of Psychopharmacology*, vol. 29, pp. 85–96, 2015.

[30] P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang, "10 years of gwas discovery: biology, function, and translation," *American Journal of Human Genetics*, vol. 101, pp. 5–22, 2017.

[31] R. J. Klein, C. Zeiss, E. Y. Chew, J. Y. Tsai, R. S. Sackler, C. Haynes, and et al., "Complement factor h polymorphism in age-related macular degeneration," *Science*, vol. 308, no. 5720, pp. 385–389, 2005.

[32] M. McGue and T. J. J. Bouchard, "Genetic and environmental influences on human behavioral differences," *Annual Review of Neuroscience*, vol. 21, pp. 1–24, 1998.

[33] H. Chial, "Mendelian genetics: Patterns of inheritance and single-gene disorders." Available: https://www.nature.com/scitable/topicpage/mendelian-genetics-patterns-of-inheritance-and-single-966/, 2008.

[34] J. F. Gusella, N. S. Wexler, P. M. Conneally, S. L. Naylor, M. A. Anderson, R. E. Tanzi, and et al., "A polymorphic dna marker genetically linked to huntington's disease," *Nature*, vol. 306, no. 5940, pp. 234–238, 1983.

[35] G. D. Schellenberg, M. A. Pericak-Vance, E. M. Wijsman, D. K. Moore, P. C. Gaskell, L. A. Yamaoka, J. L. Bebout, L. Anderson, K. A. Welsh, and C. M. Clark, "Linkage analysis of familial alzheimer disease, using chromosome 21 markers," *American Journal of Human Genetics*, vol. 48, no. 3, pp. 563–583, 1991.

[36] N. Risch and K. Merikangas, "The future of genetic studies of complex human diseases," *Science*, vol. 273, no. 5281, pp. 1516–1517, 1996.

[37] International HapMap Consortium, "The international hapmap project," *Nature*, vol. 426, no. 6968, pp. 789–796, 2003.

[38] J. N. Hirschhorn, "Genomewide association studies–illuminating biologic pathways," *New England Journal of Medicine*, vol. 360, no. 17, pp. 1699–1701, 2009.

[39] D. M. Howard, M. J. Adams, T. K. Clarke, J. D. Hafferty, J. Gibson, M. Shirali, and et al., "Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions," *Nature Neuroscience*, vol. 22, no. 3, pp. 343–352, 2019.

[40] N. R. Wray and et al., "Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression," *Nature Genetics*, vol. 50, no. 5, pp. 668–681, 2018.

[41] E. A. Stahl and et al., "Genome-wide association study identifies 30 loci associated with bipolar disorder," *Nature Genetics*, vol. 51, no. 5, pp. 793–803, 2019.

[42] P. Sklar and et al., "Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near odz4," *Nature Genetics*, vol. 43, no. 10, pp. 977–983, 2011.

[43] T. Manolio, F. Collins, N. Cox, and et al., "Finding the missing heritability of complex diseases," *Nature*, vol. 461, pp. 747–753, 2009.

[44] E. A. Boyle, Y. I. Li, and J. K. Pritchard, "An expanded view of complex traits: from polygenic to omnigenic," *Cell*, vol. 169, pp. 1177–1186, 2017.

[45] F. Dudbridge and A. Gusnanto, "Estimation of significance thresholds for genomewide association scans," *Genetic Epidemiology*, vol. 32, pp. 227–234, 2008.

[46] K. A. Frazer, S. S. Murray, N. J. Schork, and E. J. Topol, "Human genetic variation and its contribution to complex traits," *Nature Reviews Genetics*, vol. 10, pp. 241–251, 2009.

[47] H. Aschard and et al., "Inclusion of gene–gene and gene–environment interactions unlikely to dramatically improve risk prediction for complex diseases," *American Journal of Human Genetics*, vol. 90, pp. 962–972, 2012.

[48] J. McClellan and M. C. King, "Genetic heterogeneity in human disease," *Cell*, vol. 141, pp. 210–217, 2010.

[49] R. J. F. Loos and A. Janssens, "Predicting polygenic obesity using genetic information," *Cell Metabolism*, vol. 25, pp. 535–543, 2017.

[50] A. C. Janssens and C. M. van Duijn, "Genome-based prediction of common diseases: advances and prospects," *Human Molecular Genetics*, vol. 17, pp. R166–R173, 2008.

[51] M. McCarthy, G. Abecasis, L. Cardon, and et al., "Genome-wide association studies for complex traits: consensus, uncertainty and challenges," *Nature Reviews Genetics*, vol. 9, pp. 356–369, 2008.

[52] W. E. Evans and M. V. Relling, "Pharmacogenomics: Translating functional genomics into rational therapeutics," *Science*, vol. 286, no. 5439, pp. 487–491, 1999.

[53] The International SNP Map Working Group, "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms," *Nature*, vol. 409, pp. 928–933, 2001.

[54] T. A. Manolio, "Bringing genome-wide association findings into clinical use," *Nature Reviews Genetics*, vol. 14, no. 8, pp. 549–558, 2013.

[55] A. Arslan, "Imaging genetics of schizophrenia in the post-gwas era," *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 80, pp. 155–165, 2018.

[56] Z. Ramazanova, B. Matkarimov, S. Nabavi, B. Crape, and A. Zollanvari, "Snp-based prediction of schizophrenia using machine learning," *Bioinformatics Advances*, vol. 6, no. 1, p. vbaf219, 2026.

[57] S. E. Bergen and T. L. Petryshen, "Genome-wide association studies of schizophrenia: does bigger lead to better results?," *Current Opinion in Psychiatry*, vol. 25, no. 2, pp. 76–82, 2012.

[58] A. G. Cardno, E. J. Marshall, B. Coid, and et al., "Heritability estimates for psychotic disorders: The maudsley twin psychosis series," *Archives of General Psychiatry*, vol. 56, no. 2, pp. 162–168, 1999.

[59] S. Dollfus and B. Everitt, "Symptom structure in schizophrenia: two-, three- or four-factor models?," *Psychopathology*, vol. 31, no. 3, pp. 120–130, 1998.

[60] J. A. Suhr and M. B. Spitznagel, "Factor versus cluster models of schizotypal traits. i: a comparison of unselected and highly schizotypal samples," *Schizophrenia Research*, vol. 52, no. 3, pp. 231–239, 2001.

[61] J. A. Suhr and M. B. Spitznagel, "Factor versus cluster models of schizotypal traits ii: Relation to neuropsychological impairment," *Schizophrenia Research*, vol. 52, no. 3, pp. 241–250, 2001.

[62] K. T. Mueser and S. R. McGurk, "Schizophrenia," *Lancet*, vol. 363, no. 9426, pp. 2063–2072, 2004.

[63] G. Lawyer, H. Nyman, I. Agartz, and et al., "Morphological correlates to cognitive dysfunction in schizophrenia as studied with bayesian regression," *BMC Psychiatry*, vol. 6, p. 31, 2006.

[64] S. G. Schwab and D. B. Wildenauer, "Genetics of psychiatric disorders in the gwas era: an update on schizophrenia," *European Archives of Psychiatry and Clinical Neuroscience*, vol. 263, no. 2, pp. 147–154, 2013.

[65] B. K. Suarez, J. Duan, A. R. Sanders, and et al., "Genomewide linkage scan of 409 european-ancestry and african american families with schizophrenia: suggestive evidence of linkage at 8p23.3-p21.2 and 11p13.1-q14.1 in the combined sample," *American Journal of Human Genetics*, vol. 78, no. 2, pp. 315–333, 2006.

[66] T. Lencz, T. V. Morgan, M. Athanasiou, B. Dain, C. R. Reed, J. M. Kane, R. Kucherlapati, and A. K. Malhotra, "Converging evidence for a pseudoautosomal cytokine receptor gene locus in schizophrenia," *Molecular Psychiatry*, vol. 12, no. 6, pp. 572–580, 2007.

[67] H. Stefansson, D. Rujescu, S. Cichon, O. P. H. Pietiläinen, A. Ingason, S. Steinberg, and et al., "Large recurrent microdeletions associated with schizophrenia," *Nature*, vol. 455, no. 7210, pp. 232–236, 2008.

[68] The International Schizophrenia Consortium, "Rare chromosomal deletions and duplications increase risk of schizophrenia," *Nature*, vol. 455, no. 7210, pp. 237–241, 2008.

[69] M. C. O'Donovan, N. Norton, H. Williams, and et al., "Analysis of 10 independent samples provides evidence for association between schizophrenia and a snp flanking fibroblast growth factor receptor 2," *Molecular Psychiatry*, vol. 14, no. 1, pp. 30–36, 2009.

[70] L. Athanasiu, M. Mattingsdal, A. K. Kähler, A. Brown, O. Gustafsson, I. Agartz, and et al., "Gene variants associated with schizophrenia in a norwegian genome-wide study are replicated in a large european cohort," *Journal of Psychiatric Research*, vol. 44, no. 12, pp. 748–753, 2010.

[71] M. R. Salleh, "The genetics of schizophrenia," *Malaysian Journal of Medical Sciences*, vol. 11, pp. 3–11, 2004.

[72] P. V. Gejman, A. R. Sanders, and J. Duan, "The role of genetics in the etiology of schizophrenia," *Psychiatric Clinics of North America*, vol. 33, pp. 35–66, 2010.

[73] D. Bzdok and A. Meyer-Lindenberg, "Machine learning for precision psychiatry: opportunities and challenges," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 3, no. 3, pp. 223–230, 2018.

[74] D. B. Dwyer, P. Falkai, and N. Koutsouleris, "Machine learning approaches for clinical psychology and psychiatry," *Annual Review of Clinical Psychology*, vol. 14, pp. 91–118, 2018.

[75] N. Tandon and R. Tandon, "Using machine learning to explain the heterogeneity of schizophrenia. realizing the promise and avoiding the hype," *Schizophrenia Research*, vol. 214, pp. 70–75, 2019.

[76] S. Szymczak, J. M. Biernacka, H. J. Cordell, O. González-Recio, I. R. König, H. Zhang, and Y. V. Sun, "Machine learning in genome-wide association studies," *Genetic Epidemiology*, vol. 33, pp. S51–S57, 2009.

[77] A. Li and D. Meyre, "Challenges in reproducibility of genetic association studies: lessons learned from the obesity field," *International Journal of Obesity*, vol. 37, no. 4, pp. 559–567, 2013.

[78] S. Das, A. Dey, A. Pal, and N. Roy, "Applications of artificial intelligence in machine learning: Review and prospect," *International Journal of Computer Applications*, vol. 115, no. 9, pp. 31–41, 2015.

[79] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proceedings of the National Academy of Sciences*, vol. 116, no. 44, pp. 22071–22080, 2019.

[80] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[81] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[82] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning*, vol. 29, pp. 131–163, 1997.

[83] "Naive bayes classifiers." Available: https://www.geeksforgeeks.org/machine-learning/naive-bayes-classifiers/, 2024.

[84] H. Ren and Q. Guo, "Flexible learning tree augmented naïve classifier and its application," *Knowledge-Based Systems*, vol. 260, p. 110140, 2023.

[85] OpenAI, "Chatgpt (gpt-5.3)." Available: https://chat.openai.com/, 2026.

[86] A. Liaw and M. Wiener, "Classification and regression by random forest," *R News*, vol. 2, pp. 18–22, 2002.

[87] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[88] A. Venkataraman, T. J. Whitford, C. F. Westin, P. Golland, and M. Kubicki, "Whole brain resting state functional connectivity abnormalities in schizophrenia," *Schizophrenia Research*, vol. 139, no. 1–3, pp. 7–12, 2012.

[89] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 20, pp. 215–232, 1958.

[90] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[91] J. Kambeitz, L. Kambeitz-Ilankovic, S. Leucht, S. Wood, C. Davatzikos, B. Malchow, and et al., "Detecting neuroimaging biomarkers for schizophrenia: a meta-analysis of multivariate pattern recognition studies," *Neuropsychopharmacology*, vol. 40, no. 7, pp. 1742–1751, 2015.

[92] C. Toh and J. P. Brody, "A genetic risk score using human chromosomal-scale length variation can predict schizophrenia," *Scientific Reports*, vol. 11, p. 18866, 2021.

[93] R. de Filippis, E. A. Carbone, R. Gaetano, A. Bruni, V. Pugliese, C. Segura-Garcia, and P. De Fazio, "Machine learning techniques in a structural and functional mri diagnostic approach in schizophrenia: a systematic review," *Neuropsychiatric Disease and Treatment*, vol. 15, pp. 1605–1627, 2019.

[94] P. He, X. Lei, D. Yuan, Z. Zhu, and S. Huang, "Accumulation of minor alleles and risk prediction in schizophrenia," *Scientific Reports*, vol. 7, p. 11661, 2017.

[95] Y. T. Jo, S. W. Joo, S. Shon, H. Kim, Y. Kim, and J. Lee, "Diagnosing schizophrenia with network analysis and a machine learning method," *International Journal of Methods in Psychiatric Research*, vol. 29, p. e1818, 2020.

[96] I. C. Gould, A. M. Shepherd, K. R. Laurens, M. J. Cairns, V. J. Carr, and M. J. Green, "Multivariate neuroanatomical classification of cognitive subtypes in schizophrenia: a support vector machine learning approach," *NeuroImage: Clinical*, vol. 6, pp. 229–236, 2014.

[97] N. Koutsouleris, R. S. Kahn, A. M. Chekroud, S. Leucht, P. Falkai, T. Wobrock, and et al., "Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach," *Lancet Psychiatry*, vol. 3, no. 10, pp. 935–946, 2016.

[98] N. Koutsouleris, T. Wobrock, B. Guse, B. Langguth, M. Landgrebe, P. Eichhammer, and et al., "Predicting response to repetitive transcranial magnetic stimulation in patients with schizophrenia using structural magnetic resonance imaging: A multisite machine learning analysis," *Schizophrenia Bulletin*, vol. 44, no. 5, pp. 1021–1034, 2017.

[99] J. Chen, J. Wu, T. Mize, D. Shui, and X. Chen, "Prediction of schizophrenia diagnosis by integration of genetically correlated conditions and traits," *Journal of Neuroimmune Pharmacology*, vol. 13, pp. 532–540, 2018.

[100] J. Shi, D. F. Levinson, J. Duan, and et al., "Common variants on chromosome 6p22.1 are associated with schizophrenia," *Nature*, vol. 460, pp. 753–757, 2009.

[101] S. E. Bergen, C. T. O'Dushlaine, S. Ripke, and et al., "Genome-wide association study in a swedish population yields support for greater cnv and mhc involvement in schizophrenia compared with bipolar disorder," *Molecular Psychiatry*, vol. 17, pp. 880–886, 2012.

[102] T. S. Stroup, J. P. McEvoy, M. S. Swartz, and et al., "The national institute of mental health clinical antipsychotic trials of intervention effectiveness (catie) project: schizophrenia trial design and protocol development," *Schizophrenia Bulletin*, vol. 29, pp. 15–31, 2003.

[103] P. F. Sullivan, D. Lin, J.-Y. Tzeng, and et al., "Genomewide association for schizophrenia in the catie study: results of stage 1," *Molecular Psychiatry*, vol. 13, pp. 570–584, 2008.

[104] R. Buettner, M. Hirschmiller, K. Schlosser, M. Rössle, M. Fernandes, and I. J. Timm, "High-performance exclusion of schizophrenia using a novel machine learning method on eeg data," in *Proceedings of HealthCom*, 2019.

[105] A. B. Zheutlin, A. M. Chekroud, R. Polimanti, J. Gelernter, F. W. Sabb, R. M. Bilder, and et al., "Multivariate pattern analysis of genotype–phenotype relationships in schizophrenia," *Schizophrenia Bulletin*, vol. 44, no. 5, pp. 1045–1052, 2018.

[106] M. R. Arbabshirani, S. Plis, J. Sui, and V. D. Calhoun, "Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls," *NeuroImage*, vol. 145, pp. 137–165, 2017.

[107] S. Okser, T. Pahikkala, and T. Aittokallio, "Genetic variants and their interactions in disease risk prediction - machine learning and network perspectives," *BioData Mining*, vol. 6, p. 5, 2013.

[108] D. Koller and M. Sahami, "Toward optimal feature selection," in *Proceedings of the International Conference on Machine Learning*, pp. 284–292, 1996.

[109] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, pp. 273–324, 1997.

[110] S. Purcell, B. Neale, K. Todd-Brown, and et al., "Plink: Plink: a tool set for whole-genome association and population based linkage analyses," *American Journal of Human Genetics*, vol. 81, pp. 559–575, 2007.

[111] E. W. Sayers, J. Beck, E. E. Bolton, D. Bourexis, and et al., "Database resources of the national center for biotechnology information," *Nucleic Acids Research*, vol. 49, pp. D10–D17, 2021.

[112] M. D. Mailman, M. Feolo, Y. Jin, M. Kimura, and et al., "The ncbi dbgap database of genotypes and phenotypes," *Nature Genetics*, vol. 39, no. 10, pp. 1181–1186, 2007.

[113] "Genetic analysis | snp analysis | affymetrix." Available: https://docs.gdc.cancer.gov/Encyclopedia/pages/Affymetrix_SNP_6.0/, 2024.

[114] C. Anderson, F. Pettersson, G. Clarke, and et al., "Data quality control in genetic case-control association studies," *Nature Protocols*, vol. 5, pp. 1564–1573, 2010.

[115] C. Tian, P. K. Gregersen, and M. F. Seldin, "Accounting for ancestry: population substructure and genome-wide association studies," *Human Molecular Genetics*, vol. 17, pp. R143–R150, 2008.

[116] J. M. Goldstein, M. T. Tsuang, and S. V. Faraone, "Gender and schizophrenia: Implications for understanding the heterogeneity of the illness," *Psychiatry Research*, vol. 28, no. 3, pp. 243–253, 1989.

[117] G. L. Wojcik, M. Graff, K. K. Nishimura, and et al., "Genetic analyses of diverse populations improves discovery for complex traits," *Nature*, vol. 570, pp. 514–518, 2019.

[118] K. C. Barnes, "Genomewide association studies in allergy and the influence of ethnicity," *Current Opinion in Allergy & Clinical Immunology*, vol. 10, pp. 427–433, 2010.

[119] J. J. McGrath, "Variations in the incidence of schizophrenia: Data versus dogma," *Schizophrenia Bulletin*, vol. 32, pp. 195–197, 2005.

[120] A. Agresti, *Categorical Data Analysis*. Wiley, 2002.

[121] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, no. 6, pp. 1191–1253, 2003.

[122] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[123] A. Zollanvari and G. Alterovitz, "Snp by snp by environment interaction network of alcoholism," *BMC Systems Biology*, vol. 11, p. 19, 2017.

[124] "dbgap study phs000167.v1.p1." Available: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000167.v1.p1, 2024.

[125] F. Pedregosa and et al., "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[126] G. Stelzer, N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, and et al., "The genecards suite: from gene data mining to disease genome sequence analyses," *Current Protocols in Bioinformatics*, vol. 54, no. 1, pp. 1–30, 2016.

[127] D. Karolchik, R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet, D. J. Thomas, R. J. Weber, D. Haussler, and W. J. Kent, "The ucsc genome browser database," *Nucleic Acids Research*, vol. 31, no. 1, pp. 51–54, 2003.

[128] H. Sugawara, Y. Murata, and et al., "Dna methylation analyses of the candidate genes identified by a methylome-wide association study revealed common epigenetic alterations in schizophrenia and bipolar disorder," *Psychiatry and Clinical Neurosciences*, vol. 72, pp. 245–254, 2018.

[129] Y. Zhang, X. You, S. Li, and et al., "Peripheral blood leukocyte rna-seq identifies a set of genes related to abnormal psychomotor behavior characteristics in patients with schizophrenia," *Medical Science Monitor*, vol. 26, 2020.

[130] D. J. Leirer, C. O. Iyegbe, M. Di Forti, and et al., "Differential gene expression analysis in blood of first episode psychosis patients," *Schizophrenia Research*, vol. 209, pp. 88–97, 2019.

[131] G. Costain, A. C. Lionel, D. Merico, and et al., "Pathogenic rare copy number variants in community-based schizophrenia suggest a potential role for clinical microarrays," *Human Molecular Genetics*, vol. 22, pp. 4485–4501, 2013.

[132] M. G. Butler, A. B. McGuire, H. Masoud, and A. M. Manzardo, "Currently recognized genes for schizophrenia: High-resolution chromosome ideogram representation," *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, vol. 171B, no. 2, pp. 181–202, 2016.

[133] T. J. Crow, "Schizophrenia as variation in the sapiens-specific epigenetic instruction to the embryo," *Clinical Genetics*, vol. 81, pp. 319–324, 2012.

[134] E. A. H. Knauff, H. M. Blauw, P. L. Pearson, and et al., "Copy number variants on the x chromosome in women with primary ovarian insufficiency," *Fertility and Sterility*, vol. 95, pp. 1584–1588.e1, 2011.

[135] M. Wang, S.-W. Zhao, D. Wu, Y.-H. Zhang, Y.-K. Han, K. Zhao, T. Qi, Y. Liu, L.-B. Cui, and Y. Wei, "Transcriptomic and neuroimaging data integration enhances machine learning classification of schizophrenia," *Psychoradiology*, vol. 4, p. kkae005, 2024.

[136] A. Saha, S. Park, Z. W. Geem, and P. K. Singh, "Schizophrenia detection and classification: A systematic review of the last decade," *Diagnostics*, vol. 14, no. 23, p. 2698, 2024.

[137] S. Ma and J. Huang, "Penalized feature selection and classification in bioinformatics," *Briefings in Bioinformatics*, vol. 9, no. 5, pp. 392–403, 2008.

[138] T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange, "Genome-wide association analysis by lasso penalized logistic regression," *Bioinformatics*, vol. 25, no. 6, pp. 714–721, 2009.

[139] K. Ohi, C. Sumiyoshi, H. Fujino, and et al., "Genetic overlap between general cognitive function and schizophrenia: A review of cognitive gwass," *International Journal of Molecular Sciences*, vol. 19, p. 3822, 2018.

[140] G. Costain, A. C. Lionel, D. Merico, and et al., "Pathogenic rare copy number variants in community-based schizophrenia suggest a potential role for clinical microarrays," *Human Molecular Genetics*, vol. 22, pp. 4485–4501, 2013.

[141] N. Li, S. Gao, S. Wang, and et al., "Attractin participates in schizophrenia by affecting testosterone levels," *Frontiers in Cell and Developmental Biology*, vol. 9, p. 755165, 2021.

[142] Y. Kokubo, S. Padmanabhan, Y. Iwashima, K. Yamagishi, and A. Goto, "Gene and environmental interactions according to the components of lifestyle modifications in hypertension guidelines," *Environmental Health and Preventive Medicine*, vol. 24, p. 19, 2019.

[143] M. F. Feitosa, A. T. Kraja, D. I. Chasman, and et al., "Novel genetic associations for blood pressure identified via gene-alcohol interaction in up to 570k individuals across multiple ancestries," *PLOS ONE*, vol. 13, p. e0198166, 2018.

[144] K. Wang, X. Luo, and L. Zuo, "Genetic factors for alcohol dependence and schizophrenia: Common and rare variants," *Austin Journal of Drug Abuse and Addiction*, vol. 1, no. 1, p. 3, 2014.

[145] C. Johnson, T. Drgon, D. Walther, and G. R. Uhl, "Genomic regions identified by overlapping clusters of nominally-positive snps from genome-wide studies of alcohol and illegal substance dependence," *PLOS ONE*, vol. 6, p. e19210, 2011.

[146] H. M. Niu, P. Yang, H. H. Chen, and et al., "Comprehensive functional annotation of susceptibility snps prioritized 10 genes for schizophrenia," *Translational Psychiatry*, vol. 9, p. 56, 2019.

[147] K. Hirschfeldova, J. Cerny, P. Bozikova, and et al., "Evidence for the association between the intronic haplotypes of ionotropic glutamate receptors and first-episode schizophrenia," *Journal of Personalized Medicine*, vol. 11, p. 1250, 2021.

[148] Psychosis Endophenotypes International Consortium and et al., "A genome-wide association analysis of a broad psychosis phenotype identifies three loci for further investigation," *Biological Psychiatry*, vol. 75, pp. 386–397, 2014.

[149] X. Y. Wang, J. J. Lin, M. K. Lu, and et al., "Development and validation of a web-based prediction tool on minor physical anomalies for schizophrenia," *Schizophrenia*, vol. 8, p. 4, 2022.

[150] Y. Xiang, J. Wang, G. Tan, F.-X. Wu, and J. Liu, “Schizophrenia identification using multi-view graph measures of functional brain networks,” *Frontiers in Bioengineering and Biotechnology*, vol. 7, p. 479, 2020.

[151] K. T. Zondervan and L. R. Cardon, “The complex interplay among factors that influence allelic association,” *Nature Reviews Genetics*, vol. 5, pp. 89–100, 2004.

Appendices

BLANK

Appendix A

Explanatory notes

The present study applies the Cochran-Mantel-Haenszel (CMH) test to rank the SNPs using a 5-fold cross-validation training dataset. Specifically, we consider those SNPs with a Benjamini-Hochberg False Discovery Rate (BH-FDR) less than a threshold T, where T was treated as a hyperparameter of the learning rule. To this end, we investigate the performance of the learning rule for different values of T in 0.01, 0.05, 0.1. Our results demonstrate that T=0.1 consistently yielded the best performance of a BH-FDR threshold for all ethnicity-gender populations.

]

Table A.1: The estimated 5-CV AUC and accuracy for the Naïve Bayes classifier across all ethnicity-gender populations.

Model	**AA-F**		**AA-M**		**EA-F**		**EA-M**	
	AUC	**Acc**	**AUC**	**Acc**	**AUC**	**Acc**	**AUC**	**Acc**
NB	73.3%	73.2%	67.3%	66.7%	79.0%	79.6%	79.7%	71.7%

Table A.2: The estimated 5-CV AUC and accuracy for the Tree-Augmented Naïve Bayes (TAN) classifier across all ethnicity-gender populations.

Model	**AA-F**		**AA-M**		**EA-F**		**EA-M**	
	AUC	**Acc**	**AUC**	**Acc**	**AUC**	**Acc**	**AUC**	**Acc**
TAN	72.3%	73.1%	69.5%	69.8%	80.8%	80.7%	81.2%	71.6%

Table A.3: The estimated 5-CV AUC and accuracy for the Random Forest classifier with the number of base estimators in {10, 100, 1000, 10000} and maximum depth of decision trees in {2, 5, 10, 20, 50, 100} across all ethnicity-gender populations. Numbers identified by bold indicate the highest accuracy of the model.

a) AA-F population with 777 samples and 7 SNPs

N estimator	2		5		10		20		50		100	
	AUC (%)	Acc. (%)	AUC (%)	Acc. (%)	AUC (%)	Acc. (%)	AUC (%)	Acc. (%)	AUC (%)	Acc. (%)	AUC (%)	Acc. (%)
10	72.7	71.7	76.1	75.5	79.3	78.0	79.5	78.0	79.5	78.0	79.5	78.0
100	75.0	70.7	76.4	75.9	79.4	78.0	79.8	78.1	79.8	78.1	79.8	78.1
1000	74.7	70.7	76.3	76.1	79.5	78.0	79.8	78.1	79.8	78.1	79.8	78.1
10000	75.0	70.7	76.3	76.1	79.4	78.0	79.8	78.1	79.8	78.1	79.8	78.1

b) AA-M population with 776 samples and 7 SNPs

N estimator	2		5		10		20		50		100	
	AUC (%)	Acc. (%)	AUC (%)	Acc. (%)	AUC (%)	Acc. (%)	AUC (%)	Acc. (%)	AUC (%)	Acc. (%)	AUC (%)	Acc. (%)
10	67.8	68.4	70.3	70.3	70.9	70.4	70.8	70.3	70.9	71.3	70.1	70.3
100	68.7	68.8	70.3	70.9	70.3	70.9	70.3	70.9	70.3	70.9	70.3	70.9
1000	68.6	68.6	70.0	70.3	70.3	70.9	70.3	70.9	70.3	70.9	70.3	70.9
10000	68.6	68.4	70.0	70.3	70.3	70.9	70.3	70.9	70.3	70.9	70.3	70.9

c) EA-F population with 910 samples and 18 SNPs

N estimator	2		5		10		20		50		100	
	AUC (%)	Acc. (%)	AUC (%)	Acc. (%)	AUC (%)	Acc. (%)	AUC (%)	Acc. (%)	AUC (%)	Acc. (%)	AUC (%)	Acc. (%)
10	75.4	76.0	82.8	80.2	86.3	84.3	87.2	84.0	87.2	84.0	87.8	84.0
100	76.3	75.9	83.6	80.6	88.9	83.7	87.8	84.9	89.2	84.1	89.2	84.1
1000	76.6	75.8	83.8	80.8	87.0	84.3	87.8	84.0	87.8	84.0	87.8	84.0
10000	76.5	75.8	83.8	80.8	87.0	84.5	87.7	84.0	87.8	84.0	87.8	84.0

d) EA-M population with 1209 samples and 37 SNPs

N estimator	2		5		10		20		50		100	
	AUC (%)	Acc. (%)	AUC (%)	Acc. (%)	AUC (%)	Acc. (%)	AUC (%)	Acc. (%)	AUC (%)	Acc. (%)	AUC (%)	Acc. (%)
10	75.4	64.6	81.8	72.6	88.1	77.9	82.3	72.0	82.4	72.7	82.4	72.7
100	75.5	66.0	82.9	73.3	88.1	77.9	83.3	74.0	84.4	80.0	84.4	80.0
1000	76.0	66.7	83.1	74.0	89.3	80.3	83.4	74.2	83.4	74.2	83.4	74.2
10000	76.3	66.5	83.0	74.1	89.3	80.5	83.4	74.1	83.4	74.2	83.4	74.2

Table A.4: The estimated 5-CV AUC and accuracy for LR and values of λ: $\{10^{-8}, 10^{-4}, 1\}$ across all ethnicity-gender populations.

Parameter	1e-08		1e-04		1	
Population	AUC	Acc.	AUC	Acc.	AUC	Acc.
AA-F	71.9%	71.4%	71.9%	71.4%	71.9%	71.4%
AA-M	68.6%	69.7%	68.6%	69.7%	68.6%	69.7%
EA-F	79.0%	80.7%	79.0%	80.7%	79.0%	80.7%
EA-M	81.2%	71.4%	81.2%	71.4%	81.2%	71.5%

Table A.5: Selected SNPs (by CMH association test) for each ethnicity-gender population, their DNA locations, and corresponding genes for each sub-population.

AA Female

Chr	Position	SNP ID	p-value	Gene	Cytoband	Type
3	chr3:170679343	rs16835806	1.758e-12	MECOM	q26.2	intron
4	chr4:34338833	rs2125679	8.29e-10	LOC107986270	p15.1	intron
8	chr8:6430547	rs17077609	2.648e-9	MCPH1	p23.1	intron
10	chr10:50351163	rs4253197	1.02e-8	ERCC6	q11.23	intron
10	chr10:135209148	rs9629977	1.021e-6	SYCE1	q26.3	intron
14	chr14:88274566	rs4451888	9.528e-10	EML5	q31.3	intron
23	chr23:65146471	rs3903921	1.641e-9	MIR223	q12	upstream

AA Male

Chr	Position	SNP ID	p-value	Gene	Cytoband	Type
5	chr5:141817918	rs6879787	2.708e-15	SPRY4-AS1	q31.3	intron
10	chr10:50351163	rs4253197	2.259e-9	ERCC6	q11.23	intron
12	chr12:1380766	rs10773943	1.177e-5	ERC1	p13.33	intron
23	chr23:34263564	rs5973152	2.208e-8	LOC107985674	p21.1	intron
23	chr23:63725979	rs5964816	4.468e-8	YWHAZ	q11.2	upstream
23	chr23:84791655	rs6617104	7.613e-11	POF1B	q21.2	upstream
23	chr23:89022317	rs2552707	2.75e-10	TGIF2LX	q21.31	upstream

EA Female

Chr	Position	SNP ID	p-value	Gene	Cytoband	Type
2	chr2:47887055	rs3136367	1.011e-14	MSH6	p16.3	intron
3	chr3:8872402	rs11131152	3.556e-10	RAD18	p25.3	intron
3	chr3:61563456	rs9821410	1.32e-12	PTPRG	p14.2	intron
3	chr3:148380315	rs867829	1.075e-9	RPL21P71	q24	intron
5	chr5:36438538	rs16902998	3.507e-7	RNA5SP181	p13.2	upstream
5	chr5:66496146	rs17221458	8.023e-8	MAST4	q12.3	missense
5	chr5:143170695	rs224494	5.244e-8	ARHGAP26	q31.3	intron
6	chr6:45040800	rs588291	1.56e-15	SUPT3H	p21.1	intron
8	chr8:145041048	rs7816088	1.598e-20	PLEC	q24.3	downstream
9	chr9:27162125	rs10116212	3.165e-16	TEK	p21.2	intron
9	chr9:72954661	rs11142704	3.64e-8	TRPM3	q21.12	intron
9	chr9:107117818	rs10991619	5.73e-9	SLC44A1	q31.1	intron
10	chr10:27707473	rs2992009	1.378e-12	MKX	p12.1	upstream
11	chr11:78741369	rs498820	7.124e-9	TENM4	q14.1	intron
12	chr12:26581832	rs10842750	4.172e-6	ITPR2	p11.23	intron
13	chr13:51700800	rs2478955	3.842e-11	TPTE2P2	q14.3	intron
20	chr20:3535400	rs658709	4.785e-13	ATRN	p13	intron
22	chr22:45486452	rs2111399	1.612e-21	CERK	q13.31	intron

EA Male

Chr	Position	SNP ID	p-value	Gene	Cytoband	Type
2	chr2:18372735	rs16983718	4.313e-14	KCNJ3	p24.2	intron
3	chr3:8872402	rs11131152	8.922e-8	RAD18	p25.3	intron
4	chr4:62454313	rs41431344	8.592e-13	ADH1B	q13.1	intron
5	chr5:36479023	rs1692969	1.246e-7	ATP6AP1L	q14.2	downstream
5	chr5:143170695	rs224494	2.784e-9	ARHGAP26	q31.3	intron
6	chr6:45040800	rs588291	2.349e-19	SUPT3H	p21.1	intron
7	chr7:155696974	rs10230201	1.039e-5	UBEC3	q36.3	upstream
8	chr8:55583779	rs17400501	2.831e-15	SEC11B	q11.23	downstream
8	chr8:145041048	rs7816088	2.221e-21	PLEC	q24.3	downstream
9	chr9:27162125	rs10116212	8.656e-17	TEK	p21.2	intron
9	chr9:107117818	rs10991619	1.276e-14	SLC44A1	q31.1	intron
10	chr10:27707473	rs2992009	1.85e-8	MKX	p12.1	upstream
10	chr10:82445290	rs460820	1.114e-11	TENM4	q14.1	intron
12	chr12:130911503	rs17035656	1.214e-11	C12orf45	q24.33	intron
13	chr13:52112748	rs1925135	2.428e-10	HNRNPA1	q14.3	intron
14	chr14:79107235	rs1109473	5.524e-8	KIAA0743	q31.1	intron
16	chr16:20452780	rs2359518	4.341e-8	ANKRD26	p12.3	intron
20	chr20:3535400	rs658709	7.746e-9	ATRN	p13	intron
22	chr22:45486452	rs2111399	4.33e-19	CERK	q13.31	intron
23	chr23:53883679	rs6692477	3.631e-8	KAL1	p22.32	intron
23	chr23:22779838	rs7892006	2.352e-9	PTCHD1-AS	p22.11	intron
23	chr23:85431953	rs5858382	3.882e-9	IL1RAPL1	p21.3	intron
23	chr23:4499821	rs7063812	2.648e-6	TMEM47	p21.1	downstream
23	chr23:38198733	rs12012951	2.225e-6	BAIAP1	p11.4	intron
23	chr23:50762073	rs41493148	2.346e-6	SHROOM4	p11.22	intron
23	chr23:30227100	rs12014069	5.173e-9	KLF8	p11.21	intron
23	chr23:63312128	rs6524865	3.349e-8	MAMLD1	q11.2	upstream
23	chr23:2078700	rs5080003	4.339e-12	A1G4A1	q21.1	downstream
23	chr23:8611613	rs2369639	1.003e-9	DACH2	q21.31	intron
23	chr23:69022317	rs2552707	6.827e-8	TGIF2LX	q21.31	upstream
23	chr23:93360197	rs4892992	2.4e-9	TUBB4B	q21.32	intron
23	chr23:104505600	rs6523841	1.019e-8	IL1RAPL2	q22.3	intron
23	chr23:10948720	rs197027	3.696e-9	CHRD1	q25	intron
23	chr23:123434000	rs5931057	3.688e-11	COL4A5	q25	exon
23	chr23:128712780	rs5932359	4.087e-7	MTH1	q25	downstream
23	chr23:148174643	rs5980432	2.542e-24	AFF2	q28	upstream