EnSCAN: ENSEMBLE SCORING FOR PRIORITIZING CAUSATIVE VARIANTS ACROSS MULTI-PLATFORM GWAS FOR LATE-ONSET ALZHEIMER'S DISEASE

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

ONUR ERDOĞAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
THE DEPARTMENT OF MEDICAL INFORMATICS

FEBRUARY 2025

EnSCAN: ENSEMBLE SCORING FOR PRIORITIZING CAUSATIVE VARIANTS ACROSS MULTI-PLATFORM GWAS FOR LATE-ONSET ALZHEIMER'S DISEASE

Submitted by Onur Erdoğan in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Health Informatics Department, Middle East Technical University** by,

Prof. Dr. Banu Günel Kılıç
Dean, **Graduate School of Informatics** _____

Assoc. Prof. Dr. Yeşim Aydın Son
Head of Department, **Health Informatics** _____

Assoc. Prof. Dr. Yeşim Aydın Son
Supervisor, **Health Informatics Dept., METU** _____

Prof. Dr. Cem İyigün
Co-Supervisor, **Industrial Engineering Dept., METU** _____

Examining Committee Members:

Prof. Dr. Tunca Doğan
Computer Engineering Dept., Hacettepe University _____

Assoc. Prof. Dr. Yeşim Aydın Son
Health Informatics Dept., METU _____

Assist. Prof. Dr. Aybar Can Acar
Health Informatics Dept., METU _____

Assist. Prof. Dr. Burçak Otlu Sarıtaş
Health Informatics Dept., METU _____

Assoc. Prof. Dr. Öznur Taştan
Computer Science and Engineering Dept., Sabancı University _____

Date: **21.02.2025**

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Onur Erdoğan

Signature : _____

ABSTRACT

EnSCAN: ENSEMBLE SCORING FOR PRIORITIZING CAUSATIVE VARIANTS ACROSS MULTI-PLATFORM GWAS FOR LATE-ONSET ALZHEIMER'S DISEASE

Erdoğan, Onur

Ph.D., Department of Health Informatics

Supervisor: Assoc. Prof. Dr. Yeşim Aydın Son

Co-supervisor: Prof. Dr. Cem İyigün

February 2025, 87 pages

Late-onset Alzheimer's Disease (LOAD) represents a progressive and complex neurodegenerative condition prevalent among the elderly demographic. Manifesting through cognitive deterioration, including memory impairment and diminished intellectual faculties, LOAD's etiology often intertwines with traumatic brain injuries. The genetic underpinnings of Alzheimer's Disease (AD) remain elusive, impeding early and differential diagnosis of LOAD. While Genome-Wide Association Studies (GWAS) enable the examination of statistical interactions among individual genetic variants within specific loci, the traditional univariate analysis may overlook intricate relationships among these genetic elements. Conversely, machine learning (ML) algorithms prove indispensable in uncovering latent, novel, and clinically relevant patterns by accommodating nonlinear interactions among genetic variants, thereby augmenting our understanding of the genetic predisposition inherent to complex disorders. Nevertheless, conventional majority voting is inapplicable across diverse platforms due to disparate SNV's attribute configurations. Hence, a novel post-ML ensemble methodology is devised to discern significant Single Nucleotide Variants (SNVs) across multiple genotyping platforms. Introducing the EnSCAN framework, we propose a pioneering algorithm to consolidate selected variants even across distinct platforms, thereby prioritizing candidate causative loci and enhancing ML outcomes by combining prior information captured from each multimodel of each dataset. The proposed ensemble algorithm utilizes chromosomal locations of SNVs by mapping to cytogenetic bands, along with the proximities between pairs and multimodel via Random Forest (RF) validations to prioritize SNVs and candidate causative genes for Alzheimer's Disease. The

scoring method is scalable and can be applied to any multi-platform genotyping study. We present how the proposed EnSCAN scoring algorithm prioritizes the candidate causative variants related to LOAD among three GWAS datasets.

Keywords: Late-onset Alzheimer's Disease, Single Nucleotide Variation (SNV), Machine Learning, Ensemble Scoring, Causative Genes

ÖZ

EnSCAN: GEÇ BAŞLANGIÇLI ALZHEIMER HASTALIĞI (LOAD) İÇİN ÇOKLU PLATFORM GENOM ÇAPINDA İLİŞKİLENDİRME ÇALIŞMALARINDA (GWAS) HASTALIĞA NEDEN OLAN VARYANTLARI ÖNCELİKLENDİRMEK İÇİN ÇOKLU MODEL ÖNSEL BİLGİLERİN BİRLEŞTİRİLEREK SKORLANMASI

Erdoğan, Onur

Doktora, Sağlık Bilişimi Bölümü

Tez Danışmanı: Doç. Dr. Yeşim Aydın Son

Eş-Danışman: Prof. Dr. Cem İyigün

Şubat 2025, 87 sayfa

Alzheimer Hastalığı (LOAD), yaşlı bireyler arasında yaygın olarak görülen ilerleyici ve karmaşık bir nörodejeneratif hastalıktır. Hafıza bozukluğu ve azalmış entelektüel yetiler dahil olmak üzere bilişsel bozulma ile kendini gösterir. LOAD'un etiyolojisi sıklıkla travmatik beyin hasarlarıyla iç içe geçmiştir. Alzheimer Hastalığı'nın (AD) genetik temelleri belirsizliğini korumaktadır ve bu durum LOAD'un erken ve ayırt edici teşhisini engellemektedir. Genom boyutunda ilişkilendirme çalışmaları (GWAS), belirli lokuslarda bağımsız genetik varyantlar arasındaki istatistiksel etkileşimlerin incelenmesine olanak tanırken, geleneksel tek değişkenli analizler bu genetik varyantlar arasındaki doğrusal olmayan ve karmaşık ilişkileri gözden kaçırabilir. Buna karşın, makine öğrenimi (ML) algoritmaları, genetik varyantlar arasındaki doğrusal olmayan etkileşimleri dikkate alarak gizli, yenilikçi ve klinik olarak anlamlı örüntüleri ortaya çıkarmada önemli bir rol oynamaktadır. Böylece karmaşık hastalıklardaki genetik yatkınlığın anlaşılmasını artırmaktadır. Geleneksel sınıflandırma yöntemlerinden biri olan çoğunluk oy verme yöntemi, farklı genotipleme platformlarda farklı özellik yapılandırmaları nedeniyle uygulanabilir değildir. Bu nedenle, çoklu genotipleme platformları arasında anlamlı Tek Nükleotid Varyantlarını (SNV'ler) ayırt etmek için yenilikçi bir "post-ML ensemble" metodolojisi geliştirilmiştir. Bu çalışmada EnSCAN çerçevesini tanıtarak, farklı genotipleme platformlarında seçilen varyantları konsolide eden ve bu varyantları öncelikli aday lokus olarak sıralayarak her bir genotiplendirme platformundaki makine öğrenmesi sonuçlarından elde edilen önsel bilgiyi iyileştiren ve birleştiren yenilikçi bir algoritma önerilmektedir. Önerilen EnSCAN algoritması, varyatların kromozomal konumlarını

sitogenetik bantlara göre haritalama ve çiftler arasındaki yakınlıkları ile çoklu model üzerinde Rastgele Orman (RF) doğrulamalarını kullanarak SNV'leri ve Alzheimer Hastalığı için ayırt edici genleri önceliklendirmeyi sağlar. EnSCAN algoritmasındaki skorlama yöntemi ölçeklenebilir olup, herhangi bir çoklu platform genotipleme çalışmasına uygulanabilir durumdadır. Bu çalışmada EnSCAN skorlama algoritmasının, üç GWAS veri seti arasında LOAD ile ilişkili aday ayırt edici varyantlarını nasıl önceliklendirdiği sunulmaktadır.

Anahtar Sözcükler: Geç Başlangıçlı Alzheimer Hastalığı, Tek Nükleotid Varyantı, Makine Öğrenimi, Birleştirme ile Skorlama, Nedensel Genler

To My Dear Niece

ELA ÇAVDAR

ACKNOWLEDGMENTS

I appreciate my advisor who has given me great ideas and absolute support. Endless thanks to my advisor Assoc. Prof. Yeşim Aydın Son who trusted me from the beginning to the end. With the unwavering dedication of my advisor and her support in this field, my thesis proposal is also funded and completed as a TÜBİTAK 1003 project. Besides my advisor, I would like to express my heartfelt gratitude to my co-advisor, Prof. Dr. Cem İyigün, for his invaluable support and expert guidance on Machine Learning approaches. Their encouragement has been instrumental in my academic journey, inspiring me to strive for excellence and believe in my potential. I truly appreciate their dedication and the positive impact they have made on my life.

Endless thanks to my dear mother Şenay Erdoğan, my father Ahmet Mithat Erdoğan, my sister Bahar Erdoğan Çavdar, her husband Buğra Çavdar and my grandmother Servet Akdemir. Unique thanks to my 2.5-year-old niece Ela. Whenever I needed them, they were with me. Especially endless thanks to my dear mother who wakes up with me even in the early of every morning and brings me up to these days. Her being is beyond words. I would like to extend my deepest thanks to my uncle, Prof. Dr. Hasan Çelik, and my aunt Prof. Dr. Menşure Çelik, who has never wavered in his belief in me throughout my doctoral studies.

This thesis is also dedicated to my dear niece, Ela Çavdar. At the time of this work, she was just born. I hope that this thesis will inspire her in her future academic endeavors and encourage her to pursue advanced education and leadership in her field.

I also would like to extend my sincere gratitude to all my friends who continuously encouraged and supported me throughout the process of preparing and conducting my research.

I would like to express my deepest gratitude to the members of my doctoral defense committee for their invaluable contributions and support throughout this process. I extend my heartfelt thanks to my examining committee members Prof. Dr. Tunca Doğan, Assoc. Prof. Dr. Yeşim Aydın Son, Assoc. Prof. Dr. Öznur Taştan, Assist. Prof. Dr. Aybar Can Acar and Assist. Prof. Dr. Burçak Otlu Sarıtaş.

This research was supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) under the ARDEB 1003 program (Grant No: SBAG-216S468). I sincerely thank TÜBİTAK for their valuable support.

TABLE OF CONTENTS

ABSTRACT iv
ÖZ vi
DEDICATION viii
ACKNOWLEDGMENTS ix
TABLE OF CONTENTS x
LIST OF TABLES xii
LIST OF FIGURES xiii
LIST OF ABBREVIATIONS xiv
CHAPTERS
1. INTRODUCTION 1
1.1. Motivation for an Ensemble Problem 4
2. BACKGROUND 7
2.1. Alzheimer's Disease (AD) 7
2.2. Single Nucleotide Variants (SNVs) 7
2.3. Genome-Wide Association Study (GWAS) 8
2.4. Random Forest (RF) 9
2.4.1. Random Forest Implementation for Dimension Reduction 11
2.4.2. Random Forest Implementation for Modeling 12
2.5. Random Forest Tuning 12
2.6. The Out of Bag Error Estimate (OOB) 14
2.7. Measuring Feature Importance 14
2.8. Ensemble Modeling 16
3. LITERATURE REVIEWS 17
4. MATERIALS AND METHODS 23
4.1. Datasets 24

4.2. Evaluation of Random Forest In Terms of Performance Improvement in Pilot GenADA 25
4.3. Design and Implementation of Analytical Workflow for EnSCAN 25
4.3.1. Association Analysis 28
4.3.2. Feature Selection and Model Building: Individual Dataset Dimensionality Reduction and Modeling by Random Forest 32
4.3.3. Ensemble Scoring (EnSCAN) for Prioritizing Causative Variants 34
4.3.4. Variance Annotation Process 40
4.3.5. Functional Enrichment 40
5. RESULTS 43
5.1. Accuracy Results of Random Forest Algorithm with Different Train Parameters on GenADA 43
5.2. Results for GWAS Associated with LOAD 44
5.3. Model Construction and Tuning for Multi Platforms by Random Forest 45
5.4. Intersection Evaluation of Selected SNVs 50
5.5. Multi-Platform Variant Prioritization by EnSCAN Algorithm 50
5.6. Evaluation of Ensemble Scores 52
5.6.1. Functional Evaluation of All LOAD RF-RF Variants 53
5.6.2. Functional Evaluation of EnSCAN Class Score 2.31 54
5.6.3. Functional Evaluation of Union of Class Score 2.31 and 3.31 55
5.6.4. Functional Evaluation of the Union of EnSCAN Class Score 3.11, 3.21, 3.31 (all 3s): 57
5.6.5. Functional Evaluation of Class Score 3.31 58
6. DISCUSSION 61
REFERENCES 65
APPENDICES 71
APPENDIX A 71
APPENDIX B 74
APPENDIX C 79
CURRICULUM VITAE 85

LIST OF TABLES

Table 1: Databases, Case/Control Groups, and Genotype Platforms.....24
Table 2: Tuning Parameters for Multi-Platform Genome Datasets 49

LIST OF FIGURES

Figure 1: Single Nucleotide Variation 8
Figure 2: Genome-Wide Association Study 9
Figure 3: Construction [28] of Random Forest 10
Figure 4: Serial / Parallel Execution for Model Construction by Random Forest 13
Figure 5: Overview of the Modeling Methodology 23
Figure 6: Design for Model Building and EnSCAN 27
Figure 7: Comparison of GWAS p-values Relative to EnSCAN Scores 29
Figure 8: Reshape2 package: Dcast function Implementation 31
Figure 9: Merge SNV data and Phenotype for Each Dataset 32
Figure 10: Mapping to Cytogenetic Locations (Genomic Positions) on Chromosomes 35
Figure 11: Flowchart of Ensemble Scoring Algorithm (EnSCAN) 36
Figure 12: SNV is located in the middle of the window. 38
Figure 13: SNV is located at the bottom of the window. 39
Figure 14: SNV is located at the top of the window 39
Figure 15: GenADA RF Tuning Parameters 45
Figure 16: GenADA RF-RF Tuning Parameters 46
Figure 17: NCRAD RF Tuning Parameters 46
Figure 18: NCRAD RF-RF Tuning Parameters 47
Figure 19: ADNI RF Tuning Parameters 48
Figure 20: ADNI RF-RF Tuning Parameters 48
Figure 21: Number of Significant SNVs Selected with the Multimodel Approach (LOAD RF-RF) 49
Figure 22: Number of Common SNVs from Three Individual Models 50
Figure 23: Ensemble Scoring Results 51
Figure 24: Functional Enrichment Analysis of All Variants with EnSCAN Class Scores 53
Figure 25: Functional Enrichment Analysis of Variants with EnSCAN Class Score Equal to 2.31 55
Figure 26: Functional Enrichment Analysis of Union of Variants with EnSCAN Class Score 2.31 or 3.31 56
Figure 27: Functional Enrichment Analysis of Variants with EnSCAN Class Score 3.11 or 3.21 or 3.31 (3s) 57
Figure 28: Enrichment Analysis of Variants with EnSCAN Class Score Equal to 3.31 58

LIST OF ABBREVIATIONS

AD	Alzheimer's Disease
ADNI	Alzheimer's Disease Neuroimaging Initiative
AHP	Analytical Hierarchy Process
CART	Classification and Regression Tree
CPU	Central Processing Unit
dbGAP	Database of Genotypes and Phenotypes
GWAS	Genome-Wide Association Study
LD	Linkage Disequilibrium
LOAD	Late-onset Alzheimer's Disease
ML	Machine Learning
OBB	Out of Bag
RF	Random Forest
SNV (or SNP)	Single Nucleotide Variation (or Single Nucleotide Polymorphism)

CHAPTER 1

INTRODUCTION

All human beings are 99.9% identical in their genetic makeup. Genetic variations in the remaining 0.1% carry significant clues of susceptibility to a complex disease since variations in the sequence of a gene affect the protein production chain and directly the trait for the biological process. Identification of epistatic interacting effects of multiple genetic variations over particular genes plays a significant role in developing diagnosis and treatment of these complex human diseases. As both the number of genetic variations and the nature of interactions increase, such complex phenotypic traits are characterized by a high level of unpredictability.

A challenging aspect of genomic studies is to understand the biological effect of inherited genetic variations and the molecular etiology of a complex disease in DNA structure between individuals. Among genetic variations, single nucleotide variations (SNVs or SNPs) have received much attention [1] in terms of disease prediction. SNVs are the most common type of genetic variation such as insertions, deletions, and substitutions observed in the genomic DNA sequences of individuals. SNVs, as biological markers, determine personal phenotypical differences. SNVs associated with particular diseases can locate genes contributing to the pathogenetic causation and etiology behind the phenotypes of a complex genetic disease like Alzheimer's disease (AD). Discovering SNV biomarkers at different loci may improve early diagnosis accuracy and prevention of these diseases via clinical decision-making [1].

Alzheimer's disease (AD) is the most common form of dementia in aging populations where the condition is characterized by deterioration of memory and other, cognitive domains. The reasons for the late-onset variant of AD [2], which appears after age 65, are still unclear. Nowadays, the number of people affected by late-onset AD (LOAD) is set to rise dramatically. Nonetheless, a large number of single-nucleotide variants (SNVs) have been found and linked to AD [3], [4], [5], [6].

AD affects over 35 million patients worldwide, and the majority of these cases are familial showing complex genetic inheritance where interactions of multiple genetic variations affect the phenotype of patients. Even though many studies are focusing mainly on the amyloid precursors [7], [8], Presenilin [9], and Tau mechanisms [10], [11], [12] the etiology of the complex genetic form of the LOAD is unknown. This

makes the early and differential diagnosis of AD, difficult, delaying the start of therapeutics and supportive therapies; hence individual genetic variations (SNVs) and interacted variation profiles associated with diseases can be utilized as biological markers.

Identification of new AD-related genes by selecting the most suitable set of SNVs over millions of variations that would accurately represent Alzheimer's disease will provide both cost and time-effective diagnosis of the disease and an understanding of the pathological mechanisms leading to neurodegeneration. Today diagnosis of late-onset Alzheimer's disease is partially available based on clinical evaluation and imaging, and still, many patients cannot be diagnosed at the early stages of AD. SNVs which can act as biological markers by causing personal differences and locating genes that are associated with a particular complex disease, contribute to identifying the pathogenetic causation and origination behind the phenotypes of a particular disease like Alzheimer's disease (AD). Discovering SNV biomarkers at different loci may improve early diagnosis accuracy and prevention of these diseases via clinical decision-making.

In recent years, there has been a great explosion of genomic data [13]. Genomics and next-generation sequencing technology have made it possible to analyze vast amounts of data for genetic epidemiology. This occurs due to the advances in various high-throughput biotechnologies. Research centers produce high dimensional data to be able to use it for both clinical decision-making and research purposes which includes both genotyping data of different sequencing platforms such as Affymetrix or Illumina and clinical data such as diagnoses and demographics. These huge genomic data sets are information-rich and contain novel and hidden information related to a particular disease [13]. Discovering these patterns impacting cure development for a particular disease requires intelligent analysis approaches for the knowledge extraction process. One of the most challenging issues is searching and analyzing genetic risk factors that result from the interaction of multiple genes. The identification of genes that are considered markers of susceptibility to complex traits could reveal novel targets for drug development and allow for the implementation of strategies toward personalized medicine. Identifying causative SNVs among the set of biomarkers can enlighten the molecular etiology, pathogenetic changes, and the related phenotypes of complex diseases like Alzheimer's disease (AD).

Genome-wide association Studies (GWAS) represent a straightforward approach predicated on univariate hypothesis testing. GWAS evaluates individual genetic markers in isolation and determines statistical significance based on assumptions about data distribution. While this conventional method identifies novel loci associated with specific complex diseases, it has inherent limitations. Notably, GWAS does not account for genetic interactions between biomarkers, nor does it jointly present expected combinations of mutations and/or non-mutations. In other words, complicated etiologies such as SNV-SNV, and gene-gene interactions are ignored by the univariate analysis usually applied in these studies. Most importantly, although GWAS highlights the most significant correlations related to a disease, it cannot provide a predictive model that integrates these statistically significant biomarkers [14]. In this context, advanced computational techniques are required to discern patterns underlying the etiology of complex diseases. Recently, predictive models that integrate and evaluate

genetic variations alongside demographic data have been proposed to monitor the progression and facilitate the diagnosis of Alzheimer's disease. Although machine learning methods for knowledge extraction are extensively utilized in some studies [15], [16] the implementation of multimodel or hybrid [17] approaches remains limited, and the integration of information from diverse datasets is overlooked.

The association of the SNV subsets with diseases and analysis of the genotyping data with clinical findings of patients will provide practical and affordable decision-support methodologies for the prediction of diseases in the clinical domain. Such a data volume also enables new analysis methods not to make it difficult to answer research questions such as associated genes or nucleotide variations. Hence, there is a need to determine the representative SNV subset based on clinical data which provides a predictive model for susceptibility to the disease. Supervised machine learning classification learning techniques in Machine Learning on high throughput genomic data are often performed to obtain genomic disease prediction models for different groups of biological subjects and patients known as cases and controls. This approach has the highest potential for extracting knowledge from genomic datasets and to select the number of SNVs for the disease that are informative for clinical diagnosis.

Machine learning (ML) methodologies are adept at identifying significantly associated biomarkers by emphasizing variable importance and prioritization, which facilitates the development of prognostic models and decision support systems. Given the large-scale and complex nature of biological data, modern biology has increasingly incorporated advanced ML techniques [18], [19], [20]. These methods have also been adapted for use in Genome-Wide Association Studies (GWAS) as preprocessing tools, given that GWAS data can be framed as a classification problem involving thousands of individual genetic variations.

Random Forest (RF) is a supervised machine learning technique employed for classification tasks that significantly enhances classification accuracy by aggregating the votes from multiple decision trees. Random Forest methodology is a popular choice since it does not require an assumption related to the distribution of data and it is interpretable with high dimensional feature space and complex data structures and an efficient way to assess the patterns and relationships. Random Forest implementation in genomic analysis has become popular within machine learning techniques that can reveal many predictor variables and allow for generating assembled models with interactions. This method facilitates the selection of important attributes by evaluating the best splits among features. Beyond its role in constructing accurate models, RF can serve as a preprocessing step for integrating feature selection and interaction effects during the training phase. Additionally, Random Forest analysis generates metrics of variable importance [21], which can be utilized to rank and choose predictor variables. Recent research has also explored the application of Random Forest for classifying genomic data, selecting features, analyzing pathways, and predicting, associating, or detecting the underlying causes of complex diseases.

This study conducts a comprehensive meta-analysis of three authorized datasets related to late-onset Alzheimer's disease (LOAD), focusing on a comparative

assessment of two widely used microarray platforms: Affymetrix and Illumina. Through the application of multimodel machine learning techniques, we develop an innovative in-silico diagnostic model for AD. As distinct machine learning classification approaches are executed independently for each dataset, achieving optimal classification performance tailored to the unique characteristics of each dataset. Following machine learning modeling, independent models representing each multi-platform dataset are assembled to integrate information from datasets with differing metadata. A novel ensemble approach is introduced to identify significant single nucleotide variants (SNVs) across multiple microarray platforms.

The primary challenge addressed in this research is the integration of multi-step machine learning classifiers to derive a comprehensive predictive model based on prior knowledge, especially in the absence of common representative SNVs across different platforms. We mapped the SNVs identified from each dataset to cytogenetic locations to identify shared features. Specifically, SNVs are correlated with chromosomal locations, creating a new informational dimension. We utilized chromosomal locations, linkage disequilibrium (proximities between pairs), and statistical significance captured by multimodel machine learning results to add value to the prioritization by building the ensemble scoring algorithm for SNVs.

1.1. Motivation for an Ensemble Problem

The ensemble problem in supervised learning is the difficulty of merging several models or methods to enhance overall predicted performance. The concept is that while individual models could have different advantages and disadvantages, one can frequently improve accuracy and robustness by combining their predictions and extracting novel information. In machine learning applications, ensemble algorithms traditionally rely on subsets of data with similar feature sets. This study introduces a novel approach by applying ensemble methods to datasets with completely distinct feature sets. Our proposed method aims to develop an ensemble algorithm that utilizes heterogeneous metadata, effectively aggregating insights from disparate data sources. This innovative approach addresses the limitations of classical ensemble methods, such as Random Forest, which typically work with homogeneous feature sets. By incorporating the heterogeneity between datasets, this method enables the extraction of richer and more diverse insights. The contribution of this work to the literature is significant, as it broadens the scope of ensemble learning and offers a fresh perspective on evaluating algorithm performance using datasets with varied attributes, thus paving the way for more robust and versatile machine learning techniques.

This study conducts a meta-analysis of three distinct, authorized access high throughput datasets related to late-onset Alzheimer’s disease (LOAD). A comparative analysis is performed across two prevalent microarray platforms: Affymetrix and Illumina. The study employs multimodel machine learning techniques to construct an in-silico diagnostic model for Alzheimer's Disease. This model construction process is notably innovative as it treats classification approaches for each dataset separately,

thereby achieving optimal classification performance independently for each dataset. After developing machine learning (ML) models, the challenge is to integrate these distinct models from each multi-platform dataset, given that the metadata and attributes vary across platforms, precluding traditional majority voting methods. To address this, a novel ensemble approach is introduced to identify significant Single Nucleotide Variants (SNVs) across multiple microarray platforms.

The objective of this study is to integrate multi-step machine learning classifiers into a cohesive ensemble model to derive a generalized prediction model from prior knowledge, where common representative SNVs are absent across platforms. SNVs identified from each dataset are mapped to cytogenetic locations to uncover commonalities as a new informational dimension. We leveraged chromosomal locations, linkage disequilibrium (proximities between pairs), and multi-platform database characterizations (appeared in databases), along with statistical significance obtained from RF-RF modeling results as a robust validation, to enhance prioritization by considering variants that appear repetitively in the model during the scoring of prioritized SNVs.

There are several goals of this study which can be listed as follows:

The first goal is to determine and apply preliminary analysis to each genotyping dataset separately as a dimension reduction process. We expect that the Random Forest (RF) algorithm may be conducted as a preprocessing step for the selection of statistically significant features. Thereby, the accuracy rate of the model can be compared with our previous study [22] considering the same data set. If there is a significant increase in model accuracy, we will use RF as a preprocessing step for each dataset.

The second and crucial goal is to integrate the selected best distinct models of the multimodel approach and merge all prediction results into one model using an ensemble scoring algorithm for the detection of Alzheimer's disease and the representation of the molecular basis of the disease more comprehensively.

Finally, optimization of SNVs by assessing the pathways via a graphical network model of AD is utilized.

The proposed significant variants identified by this ensemble model are expected to facilitate new research aimed at the early diagnosis and prediction of LOAD and the development of novel therapeutics. Early diagnosis and treatment of such high-risk diseases are anticipated to yield substantial social and economic benefits.

CHAPTER 2

BACKGROUND

2.1. Alzheimer's Disease (AD)

The most prevalent type of dementia, Alzheimer's disease (AD), is a progressive neurodegenerative and incurable genetic brain disorder that is characterized by a slowly progressing loss of cognitive functions. Over time, progression has an impact on mental processes including behavior, memory, and thought. The disease results in severe memory failure, confusion, poor judgment, and finally death if full-time medical care is not given by early diagnosis of the disease.

The two types of Alzheimer's disease are commonly distinguished. The rarer form is early onset which especially appears before 65 years of age while the late onset symptoms occur after 65 years of age and are known as heritable genetically complex mechanism. Early onset AD starts before the age of 65. The autosomal dominant mode of inheritance is evident in this uncommon case of familial AD. Its fundamental cause is found to be based on differences in one of the three genes known as Tau, Presenilin, or Amyloid Precursor. Late-onset AD is the more prevalent type (LOAD). It accounts for 95% of all cases of AD.

Despite being inherited, its fundamental cause is a complicated system involving gene interaction. The apolipoprotein (APOE) locus has been identified as one risk factor [3], [6]. Many additional susceptibility loci for AD have been found by recent disease association studies [23]. These loci are widely distributed in the population but have relatively little effect on the risk of AD. Because of this, a significant portion of the complexity and heritability of AD are yet unknown for the genes that cause the disease.

2.2. Single Nucleotide Variants (SNVs)

The human genome comprises approximately 3.2 billion nucleotides, with 99.9% of the genome being similar across all individuals [24]. The differences observed among individuals are attributed to just 0.1% of the genome sequence, which varies approximately every 300 to 1,000 base pairs. These variations in the genome may be the results of repetitive elements, mutations, or single nucleotide variants.

Figure 1: Single Nucleotide Variation

Single Nucleotide Variants (SNVs) refer to genetic variations that involve a change in a single nucleotide base pair within the DNA sequence as shown in Figure 1. SNVs can manifest in various forms, such as substitutions where one nucleotide is replaced by another and may occur in both coding and non-coding regions of the genome. In coding regions, SNVs can lead to changes in the amino acid sequence of proteins, potentially impacting protein function and contributing to phenotypic diversity or susceptibility to disease. In non-coding regions, SNVs might affect regulatory elements, influencing gene expression levels.

SNVs are essential for comprehending genetic predispositions to diseases, individual reactions to therapies, and the underlying mechanisms of numerous biological processes because of their prevalence and possible functional implications. Genetic research and personalized medicine have benefited greatly from the enormous improvements in SNV detection and interpretation brought about by advances in genomic technologies and analytical techniques.

As multiple interactive SNVs on genes may simultaneously be responsible for complex disorders [24], linking and mapping SNVs to complicated disease occurrences is crucial. Due to the abundance of SNVs found in the human genome, association studies can be used to connect genetic differences to phenotypic changes.

SNVs also contribute to identifying the genetic etiology of a particular disease like Alzheimer's disease (AD). Discovering SNV biomarkers at different loci may improve early diagnosis and treatment of these diseases. In addition to this, the possibility of determining individuals at increased genetic risk for Alzheimer's disease (AD) may streamline biomarker, and drug trials via pharmacogenomics studies and aid clinical decision-making for diagnosing the disease at the early steps.

As far as genotyping millions of SNVs is highly expensive, it is required to obtain a representative subset of SNVs to accurately represent the disease associations.

2.3. Genome-Wide Association Study (GWAS)

A genetic association study aims to determine statistical relations between genotypes (genetic variants) and phenotypes (traits or disease states) and thus to find genetic risk factors related to a particular disease. Straightforward approaches such as Genome-Wide Association Studies (GWAS) are based on testing univariate hypotheses. GWAS does not assess the potential association of each genetic marker, and it gives statistical significance based on statistical assumptions of data distribution. The main idea behind

a GWAS is to genotype a collection of cases (affected) and controls (unaffected) individuals for a high dimension of SNVs spread over the entire genome. Genetic association studies assess correlations between genetic differences and conditions to a particular disease in a population statistically as illustrated in Figure 2. Genomic regions that show statistical differences in genotype frequencies among cases and controls in terms of phenotype of interest are then detected to use this dense collection of SNVs as candidate disease susceptibility genes.

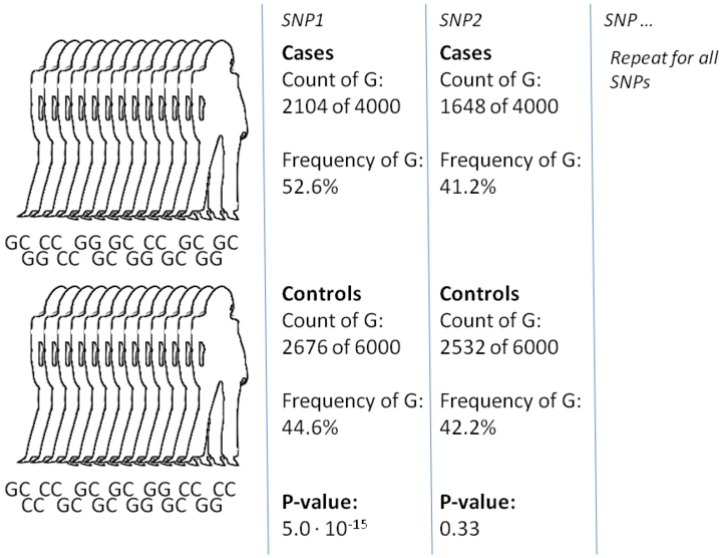


Figure 2: Genome-Wide Association Study

Although this standard method gives information about the novel loci for a particular complex disease, it has some limitations. GWAS does not consider genetic interactions of each biomarker which also means expected specific combinations of variations are not jointly presented. Most importantly, although it reveals top-ranked significant correlations associated with a particular disease, the method does not provide a prediction model to reveal the statistically significant biomarkers [14].

With today's high throughput genotyping array or whole genome sequencing technologies, millions of SNV changes can be evaluated in a single study, and areas that are altered more frequently in case groups compared to controls can be linked to the condition by GWAS. To determine complex disease-associated SNVs that change the protein production process, further analysis can be derived using the result of genome-wide association studies.

2.4. Random Forest (RF)

RF term first came from a random decision forest that was proposed by Tin Kam Ho in 1995. The method is enhanced by combining the “bagging” idea and a random

selection of features. Random Forest was proposed by Leo Breiman in 2001 [25] which is the extension of his previous work on classification and regression trees [26].

RF bagged CART trees by randomizing and selecting samples with replacements in the training set for obtaining a full tree. The application starts with selecting a bootstrap sample of the data set for growing the classifier. Within each bootstrap sample, nearly 1 out of 3 of the original data will be unselected, which is called an out-of-bag (OOB) sample. These left-out examples can be used to measure estimation accuracy [27]. Splitting variables are selected randomly. Optimal splitting criteria are searched over. This is repeated at each node until an unpruned CART is formed. The data that is not in the bootstrap sample is run down the tree to obtain the error rate and measures of variable importance. This process is iterated until the full forest is grown.

Both random sampling and ensemble strategies utilized in RF enable it to succeed in accurate predictions as well as better generalizations. This generalization property comes from the bagging scheme which provides improving the generalization. The main idea behind ensemble methods is that a group of "weak learners" can come together to form a "strong learner".

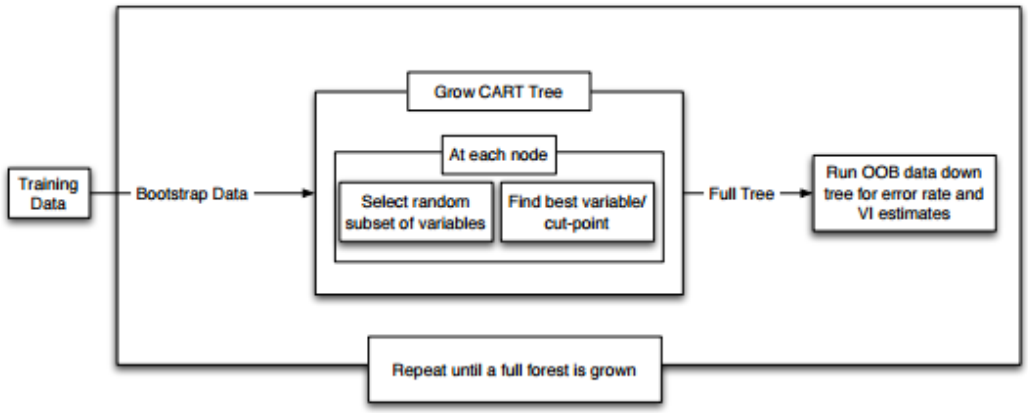


Figure 3: Construction [28] of Random Forest

Figure 3 describes how such an RF system is learned from training samples, for some number of trees T :

- Sample N cases randomly with replacement to create a subset of the data for obtaining a training set. The subset should be about 66% of the total set. When the training set for the current tree is drawn by sampling with replacement, about one-third of the cases are left out of the sample. This OOB (out-of-bag) data is used to get a running unbiased estimate of the classification error as trees are added to the forest. It is also used to get estimates of variable importance.

- At each node:
    - For some number *m*, *m* predictor variables are selected randomly from all the predictor variables.
    - The predictor variable that provides the best split according to some objective function such as the Gini index or other impurity measures is used to do a binary split on that node.
    - At the next node, another *m* variables are chosen randomly from all predictor variables, and the process is repeated until the full tree is constructed.

Depending upon the value of *m*, there are three slightly different systems:

- Random splitter selection: $m = 1$
- Breiman's bagger: $m = total\ number\ of\ predictor\ variables$
- Random forest: $m << number\ of\ predictor\ variables$. Brieman suggests three possible values form $= ½\sqrt{m}, \sqrt{m}$, and $2\sqrt{m}$.

RF trees are grown to maximal depth. There is no pruning procedure within the training process. Such a fully grown tree will be fairly unbiased but will be able to have a high variance. The main purpose of getting forest is to reduce variance by bagging and randomization. Randomization is done within the growing process. Before finding the optimal splitting criteria in each iteration, a subset of variables ($mtry << p$) is selected. *Mtry* is also considered as a tuning parameter. As *mtry* is small, the correlation between trees is less. However, small *mtry* will lead to more biased trees. If *mtry* is chosen greater, potential variance reduction via bagging is possible. This choice is the trade-off between bias and variance.

In a tree construction with ID3 or C4.5, one decision tree is built by the divide and conquer algorithm, on the other hand number of decision trees are built during the implementation of the random forest process. Random Forest uses Gini Index-based impurity measures for building decision trees. A vote from each of the decision trees is taken into consideration in deciding the final class label of a subject, which is called the ensemble process. This is a democratic selection process. Its key benefits are avoiding overfitting by reducing variance and predicting in an accurate way than other algorithms do. Moreover, they are widely preferred as runtimes are quite fast, and they can deal with unbalanced and missing data.

2.4.1. Random Forest Implementation for Dimension Reduction

Random Forest, a robust ensemble learning technique, is often employed for dimension reduction in high-dimensional datasets. This method constructs multiple decision trees during training and aggregates their outputs to enhance predictive accuracy and control overfitting. One of the key advantages of Random Forest in dimension reduction lies in its ability to provide estimates of variable importance

through measures such as the mean decrease in impurity or mean decrease in accuracy. By evaluating these importance scores, researchers can identify and select the most influential features while discarding less relevant ones. This process not only simplifies the model but also improves its interpretability and computational efficiency. Additionally, Random Forest's inherent capacity to handle large and complex datasets makes it particularly effective in reducing dimensionality without losing critical information. The resulting reduced feature set retains the most relevant variables, facilitating more manageable and insightful analyses while preserving the integrity of the underlying data patterns.

2.4.2. Random Forest Implementation for Modeling

Determining genetic variations associated with complex diseases in high-dimensional data is a prominent purpose of GWAS. Nonetheless, interactions between genetic locations are ignored by the univariate analysis. Random Forest (RF) is a very popular machine learning method that can deal with large amounts of data and allow complex models considering interactions between variables. RF has been proposed for genomic data analysis in recent studies [14], [19]. As RF prioritizes variables based on importance measured by the ranking approach, it has been arousing interest for filtering purposes in addition to modeling. In high dimensional genomic data, variable importance measures of RF can be used for screening or filtering by selecting top-ranking genetic variations for association study since RF reflects significant interactions compared to univariate approaches. RF methodology is not designed to explicitly test for the presence of interactions with a hypothesis test of significance. Conversely, variable importance measures of RF are expected to assess the overall impact of a variable on prediction. The performance of RF over genomic data has been investigated and RF variable importance measures have been outperformed [29] Fisher's Exact Test which is the measurement of genome-wide association studies.

2.5. Random Forest Tuning

To discover the best performance that Random Forest generates, parameters should be tuned instead of using default ones.

The *caret* function is used in R as a manual grid search. The number of split variables (*mtry*) is defined as a vector of possible values of predictor variables.

Since *mtry* values are the functions of predictors, 4 different values are rounded as below as an example. Consequently, a grid vector for *mtry* values is formed as $c(mtry_1,\ mtry_2,\ mtry_3,\ mtry_4)$ in R. *mtry* calculations are given below.

$$mtry_1 = round\left(log_2(number\ of\ SNVs)\right)$$

$$mtry_2 = round\left(\frac{1}{2} * (number\ of\ SNVs)\right)$$

$$mtry_3 = round\left(\sqrt{number\ of\ SNVs}\right)$$

$$mtry_4 = round\left(2 * \sqrt{number\ of\ SNVs}\right)$$

Another important parameter is the number of trees (*ntree*) to be generated in each forest. *Ntree* brings computational effort but as it *ntree* increases, the accuracy of the model may be increased. To estimate the tuning parameters, sequential *ntree* values are trained. A vector with the values $c\left(1, seq(50, maxvalue, by = 50)\right)$ is given sequentially to the training function of Random Forest in R.

Doing sequential constructions may take much time. To decrease execution time, tasks can be parallelized in R. Building an ensemble of classification and regression trees in a random forest algorithm (randomized decision trees) can be built independently from one another and in parallel across tens of CPUs and can be combined to form the random forest.

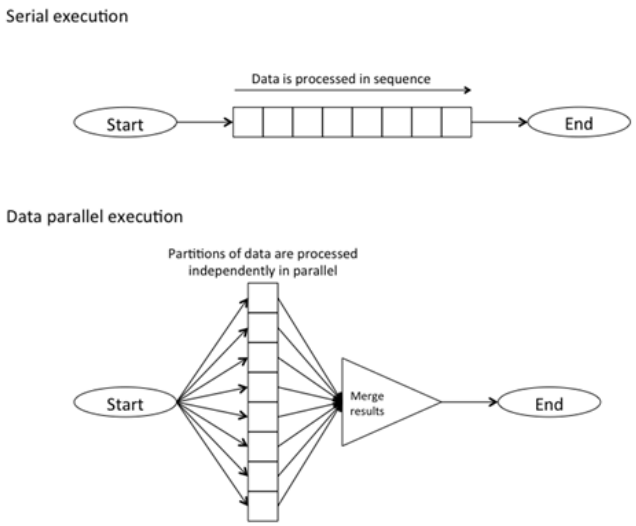


Figure 4: Serial / Parallel Execution for Model Construction by Random Forest

On one processor, both random classification trees are constructed and merged for the majority voting. Hence, assuming 50% of the task is construction and 50% of the task is combining. For a particular Random Forest, classification trees can be generated in parallel since trees are independent of each other. But, at the end of the process, all the trees must be merged. In another word, all classification trees are generated in a

parallel manner, and trees are merged at the end of the process to determine overall OOB error as illustrated in Figure 4.

2.6. The Out of Bag Error Estimate (OOB)

In a random forest, each tree is constructed using a different bootstrap sample from the original dataset. Approximately about one-third of the dataset is left out and not used in the construction of the tree.

OOB data is traversed down the tree to get a class label. In this way, a test set classification is obtained for each sample in about one-third of the dataset. At the end of the run, the class label of the majority voting is given to the sample in OOB as an observed class. The proportion of times that the observed class value is not equal to the true class of all averaged over all cases can be calculated as an OOB error estimate for the classification.

2.7. Measuring Feature Importance

The main aim of feature selection from large amounts of data is to avoid overfitting improve model performance and gain a deeper insight into the underlying processes that generated the data.

One of the machine learning strategies, the Random Forest Algorithm, directly performs feature selection while classification rules are built in the learning phase. In bioinformatics, it is widely used to select a subset of genetic markers relevant to the prediction of a specific disease. Feature importance is used to rank features based on significance.

Measuring feature importance in Random Forest involves assessing the contribution of each feature to the predictive power of the model. Mean Decrease Impurity (MDI), Mean Decrease Accuracy (MDA), and Permutation Importance are the common methods used for feature importance. As permutation importance provides a more direct measure of the feature's contribution to model accuracy, we selected the permutation-based feature importance method.

Permutation-Based Feature (Variable) Importance

RF permutation importance is an important feature selection measure. RF permutation importance is measured by permuting randomly the feature variables and computing the increase in the out-of-bag estimate of the accuracy loss. Specifically, to measure the importance of a feature in RF trees, the value of this feature is randomly shuffled in the OOB samples.

RF permutation importance measure encompasses the impact of each variable individually while considering multivariate interactions with other features at the same

time. It adopts an intuitive permutation strategy, and it is implemented more frequently than Mean Decrease Impurity (MDI) or Mean Decrease Accuracy (MDA) in general.

The permutation process involves the following steps for each feature:

- The values of the feature are shuffled across all samples within the validation dataset. This procedure effectively disrupts any inherent relationship between the feature and the target variable.

- Subsequently, the trained model is employed to generate predictions based on this permuted dataset, and the performance metric is recalculated to assess the impact of the permutation on model accuracy.

- Calculate the difference in performance between the original model and the permuted model. A significant drop indicates that the feature is important for predictions.

The permutation process is repeated multiple times (e.g., 100 or 1000 times) for each feature. This allows for a distribution of performance metrics to be collected. For each feature, a t-test is implemented, and the p-value is calculated to determine if the feature plays a significant role in accuracy.

The t-test:

The t-test is used to compare means and assess if there is a significant difference between groups. In the context of feature importance, it could be used to compare the importance scores of features against a baseline or null distribution.

If we have a distribution of importance scores obtained through permutation or bootstrapping, a t-test could be employed to assess whether the observed importance score significantly differs from this distribution. This test provides a p-value indicating the likelihood of observing such an importance score by chance.

The p-value Calculation:

In statistical analysis, p-values are used to determine whether the observed effect (such as the importance of a feature) is statistically significant.

For feature importance metrics, one might use permutation tests or bootstrapping to estimate the distribution of importance scores under the null hypothesis (i.e., no actual relationship between the feature and the response variable). The p-value is then calculated based on how extreme the observed importance score is relative to this null distribution.

In summary, RF provides feature importance scores based on impurity reduction or accuracy decrease. While it does not directly calculate p-values or perform t-tests, these statistical methods can be used in complementary analyses to assess the

significance of the feature importance scores. Specifically, permutation tests or bootstrapping methods can be used to estimate null distributions and calculate p-values to understand whether the observed importance scores are statistically significant.

2.8. Ensemble Modeling

Model ensembling is the process of combining two or more related but different analytical models. The ensemble process synthesizes the results into a single score or spread to improve the accuracy of predictive analytics and machine learning applications.

In predictive modeling, a single model can have biases, high variability, or outright inaccuracies that affect the reliability of its final findings. Using specific modeling can present similar drawbacks. Integrating different models for the same classification problem can reduce the effects of those limitations and provide better information and a higher accuracy rate for the classification of test samples.

Well-known and widely used ensemble modeling is a random forest model since it integrates multiple trees for selecting the important predictors. In other words, it is designed to predict outcomes based on different variables and rules from different trees in the forest. A random forest model blends decision trees that may analyze different sample data, evaluate different factors, or weigh common variables differently. The results of the various decision trees are then either converted into a simple average or aggregated through further weighting.

Different machine learning algorithms for the same classification problem may also be ensembled to form a better and more robust prediction model. With this approach, all the variables will be weighted according to their importance. Consequently, the final predictive model will include all important variables to diagnose the disease. In the literature, there is no similar approach implemented from different data sources and different classifiers. But Random Forest implementation and model assembling have been widely seen [14], [30], [19], [31], [32].

CHAPTER 3

LITERATURE REVIEWS

Genomic studies generate vast amounts of data, with thousands of genetic variations identified among individuals. By analyzing single nucleotide variations within the genome, researchers can correlate [33], and [34] genetic information with phenotypic traits, helping to identify nucleotide variations associated with specific diseases. These analyses often involve case-control studies aimed at discovering genetic variations that contribute to disease or serve as biomarkers. Given the sheer volume of data, extracting meaningful patterns that could aid in disease treatment development requires advanced analytical methods. A significant challenge is identifying genetic risk factors that arise from the interaction of multiple genes. Uncovering genes linked to susceptibility to complex traits can reveal new drug targets and support the advancement of personalized medicine strategies [35]. Recent advancements in genome-wide association studies (GWAS) have significantly enhanced the ability to identify new variants and genes associated with complex diseases. By pinpointing statistically significant genetic variations, the underlying mechanisms driving various phenotypes can be better understood. This understanding not only improves diagnostic precision but also paves the way for future disease prevention strategies. Several studies have been undertaken to explore genetic risk factors related to Alzheimer's disease.

Most GWAS studies [6], [36] confirmed the acceptance of the role of APOE as a genetic risk factor for late-onset AD. A genome-wide association study has resulted in the identification of 20 novel genetic risk loci in addition to APOE ε4 [10], which have been associated with a range of biological processes including lipid metabolism, immune response, and synaptic processes. The other signals from another GWAS are mostly observed in GAB2 (GRB2-associated binding protein 2), followed by less consistently replicated signals in the galanin-like peptide (GALP), piggyBac transposable element derived 1 (PGBD1), tyrosine kinase, non-receptor 1 (TNK1). Furthermore, consistent replication has been recently announced for CLU (clusterin, also known as apolipoprotein J) [37]. However, it is still not clear whether GWAS can help to reveal the underlying genetic risk factors for AD. The clinical utility of these findings for today's patients, e.g. in risk prediction, is still limited. Most of these risk factors exert only minor effects on susceptibility to AD. Nevertheless, there may already be a significant purpose for genetic risk profiling. AD is a multifactorial and

heterogeneous condition. Genetic risk profiling can be used for individual molecular subclassification, by determining the pathways enriched for risk alleles within an individual. This may increase the efficiency and outcome of clinical trials and biomarker studies. As genome-wide association studies find statistically significant variations through the genome, considering the biological impact of each variation on the susceptibility of AD will provide insight into the etiology of the disease. A study dealt with the determination of the genes and biological pathways related to disease considering prioritized genetic variations based on an Analytical Hierarchical Process [38].

In our previous research [22] using the GenADA dataset obtained from dbGAP in the context of a Master of Science thesis, we conducted a computational study on Alzheimer's disease using single nucleotide variations. We applied decision tree methodology to predict AD. Initially, we identified statistically significant SNVs using PLINK and calculated their correlations. Subsequently, we prioritized these SNVs using p-values and biological data, such as the functional effects of the SNVs, with the help of METU SNP, which laid the groundwork for selecting a subset of SNVs. For scoring and prioritizing the SNVs, we employed the Analytic Hierarchy Process (AHP) [38] to reduce dimensionality as a preprocessing step. During the modeling phase, we identified predictors through attribute selection criteria and constructed two decision trees: ID3 and C4.5. The ID3 tree was built using 958 SNVs, while the C4.5 tree incorporated both these SNVs and additional clinical information about individuals. In the ID3 tree, attribute selection criteria refined the dataset to 38 SNVs. In contrast, the C4.5 tree, which considered clinical features in addition to SNVs, selected 27 SNVs based on the same criteria. The results indicated that clinical information has a limited impact on diagnosing Alzheimer's disease when comparing the accuracy of the two decision trees. Specifically, the decision tree built solely with SNV data achieved an accuracy rate of 56.08%, whereas the decision tree that included both SNV and clinical data had a slightly lower accuracy rate of 55.07%. This preliminary study suggests that Alzheimer's disease diagnosis can be effectively conducted using only genetic information, underscoring a significant genetic component in the development of the disease.

Univariate statistical analyses, which are commonly used in genome-wide association studies (GWAS), often overlook complex interactions like gene-gene or SNV-SNV interactions. To address this, advanced techniques have been developed to detect nonlinear patterns and uncover conditional dependencies that influence phenotypes. Machine learning methods are employed to identify new patterns within extensive datasets. Various machine learning algorithms have been applied to genomic data to handle and analyze large volumes of genetic information. These include ensemble methods such as Random Forests [31], [30], [39], [40], Bayesian approaches [41] [42], [43], Decision Trees [22], Support Vector Machines [17], and various Regression techniques [44]. In the literature, there is no existing study that combines prior information from different data sources to establish a predictive model for preventive diagnosis in terms of a specific disease.

One study [39] as a future selection method, explained that the application of the Random Forests method elucidates the interaction between important genetic locus in addition to finding main effects as univariate analysis does. However, one consideration was that default tree construction parameters were not appropriate for large GWA datasets. Hence, optimization of parameters is needed. Furthermore, strong effects can be removed or pruning based on linkage disequilibrium can be handled for a better determination of significant genetic variations in Random Forests. Results obtained new interesting candidate genes such as MPHOSPH9, CTNNA3, PHACTR2, and IL7 for further follow-up studies.

In another study [42] based on a probabilistic approach, a model-averaged naive Bayes method was proposed with the use of purpose for ranking features or predicting the outcomes. In this approach, averages of predictions of models (BMA) which contain different sets of variables were weighted by the posterior probability of each model. BMA is based on the notation of averaging over a set of all possible models (2n) constructed by naïve Bayesian. Based on the proposed model 10 SNVs with the highest posterior probability of an association with LOAD were revealed such as rs429358 in APOE2, rs4420638 in APOC, rs7412 in APOE, rs10824310 in PRKG and located in unknown genes rs3732443, rs16974268, rs13414200, rs3905173, rs16939136 and rs9453276. These SNVs may be input for further investigation studies.

Identifying SNV biomarkers at various loci can facilitate early detection and improved treatment of late-onset Alzheimer's disease (LOAD). A study applied Bayesian Networks and their deviations [41] to a LOAD dataset, as this method effectively presents interactions between genetic variations and visualizes dependent connections. After performing GWAS for dimension reduction, a network was built, highlighting the top ten genes based on SNV locations. Significant SNVs identified included rs7530069, rs113464261, rs114506298, rs73504429, rs7929589, rs76306710, and rs668134. The results showed that SNVs selected using Markov blanket-based methods were strongly associated with AD and outperformed naive Bayes and tree-augmented naive Bayes models. The minimal augmented Markov Blanket-based method achieved the highest accuracy at 66.13% and a sensitivity of 88.87%.

To identify clustered SNVs associated with late-onset Alzheimer's disease (LOAD) around specific pathways, a study employed two distinct machine learning approaches [31]. Both approaches utilized logistic regression as a preprocessing step for dimensionality reduction, selecting minimum p-values that independently correlated with the disease. This process was a single-locus analysis. In the first approach, the selected SNVs were analyzed using the Random Forest algorithm to address multi-locus dependencies. This method identified 199 multi-locus associated SNVs linked to genes involved in calcium signaling, cell adhesion, endocytosis, immune response, and synaptic functions. The second approach combined linkage disequilibrium with logistic regression to filter loci before applying Random Forest. This approach revealed nineteen genetic variations in the ST5, TRPC1, ATG10, ANO3, NDUFA12, and NISCH genes, as well as in the APOE and GAB2 SNVs.

Whole genome variants provide valuable information for predicting human disease risk and assessing animal and plant breeding values. Despite the development of numerous advanced statistical methods to improve predictive accuracy, no single method has emerged as superior due to its varying strengths and limitations. A study [45] introduces an Ensemble Learning Method for the Prediction of Genetic Values (ELPGV), which combines several established approaches, including GBLUP, BayesA, BayesB, and BayesCπ, to enhance prediction accuracy. Validation with various well-known and simulated datasets demonstrated that ELPGV significantly improves predictive performance compared to individual methods, with p-values ranging from 4.853E−118 to 9.640E−20 for the WTCCC dataset. ELPGV effectively leverages the strengths of multiple methods, offering higher accuracy, ease of implementation, and speed, while requiring no genotype data, making it a promising tool for broad applications in genetic prediction.

Understanding non-linear genetic interactions is crucial for elucidating the genetic basis of complex diseases. While Genome-Wide Association Studies (GWAS) have identified significant associations [31], [36] between SNVs and various complex diseases, they often fail to capture the intricate genetic interactions involved. Another study [46] introduces a novel prioritization workflow combining two-step Random Forest (RF) modeling with entropy-based analysis to address this gap. Specifically, the workflow integrates PLINK filtering with RF-RF modeling, followed by an entropy-based three-way interaction information (3WII) method to uncover hidden patterns of non-linear genotype relationships in late-onset Alzheimer's disease (LOAD). The methodology involved developing three models using different datasets, which successfully detected third-order interactions often overlooked in traditional epistatic studies. The 3WII analysis, combined with PLINK filtering and RF-RF modeling, led to a refined SNV set across datasets, with specific SNVs showing associations with Alzheimer's Disease and other neurological disorders. The identified SNVs were linked to genes involved in crucial biological pathways, such as calcium ion binding and extracellular matrix interactions, which are relevant to LOAD. These findings suggest that the entropy-based 3WII approach is effective in revealing complex genetic interactions and offers potential candidate biomarkers for LOAD diagnosis.

Genome-wide association studies (GWAS) have revealed numerous variants linked to complex traits, yet pinpointing the causal genes remains a significant challenge. Another study [47] introduces an open resource that systematically fine-maps and prioritizes genes across 133.441 published human GWAS loci. By integrating genetic data from sources such as the GWAS Catalog and UK Biobank with transcriptomic, proteomic, and epigenomic information—including disease-disease and disease-molecular trait colocalization across 92 cell types and tissues—the resource identifies 729 loci with a single-coding causal variant and colocalized gene. A machine learning model was developed using the fine-mapped genetic and functional genomics data, as well as 445 curated GWAS loci, to more accurately distinguish causal genes from adjacent genes compared to traditional distance-based models. The prioritized genes demonstrated significant enrichment for known drug targets, with an odds ratio of 8.1 (95% confidence interval: 5.7 to 11.5).

Proteins that interact often have similar functions and influence related organismal traits. Network-based approaches can enhance the identification of candidate genes associated with traits discovered in genome-wide association studies (GWAS). Another study [48], applied a network-based expansion method to trait-associated genes across 1,002 human traits, demonstrating its effectiveness in uncovering known disease genes and drug targets. By analyzing the similarity of network expansion scores, groups of traits that likely share common genetic and biological processes are identified. 73 pleiotropic gene modules associated with multiple traits, which were notably enriched for genes involved in protein ubiquitination and RNA processing. Unlike gene deletion studies, this approach highlights multicellular processes and facilitates the identification of genes with potential relevance to disease. For instance, gene modules to known pathogenic variants are mapped, offering insights into drug repurposing opportunities. Additionally, network expansion scores to GWAS loci for inflammatory bowel disease, identifying relevant genes with substantial functional and genetic support is applied.

CHAPTER 4

MATERIALS AND METHODS

In this study, we gathered SNVs from three distinct GWAS datasets: ADNI, GenADA, and NCRAD. Each dataset contains SNV data coming from various genotyping platforms. We initiated the process with a quality control check on the genotyping and phenotype data, followed by a filtering procedure. Subsequently, PLINK analysis was employed to detect SNVs associated with LOAD in each dataset, and we prioritized these SNVs based on significance on accuracy using the RF-RF approach. Then we arranged the data considering the proximity of the SNVs within the same chromosomal region and the ensemble scoring algorithm that we developed calculated the prioritization scores for each SNV. Figure 5 briefly explains the roadmap for ensemble scoring.

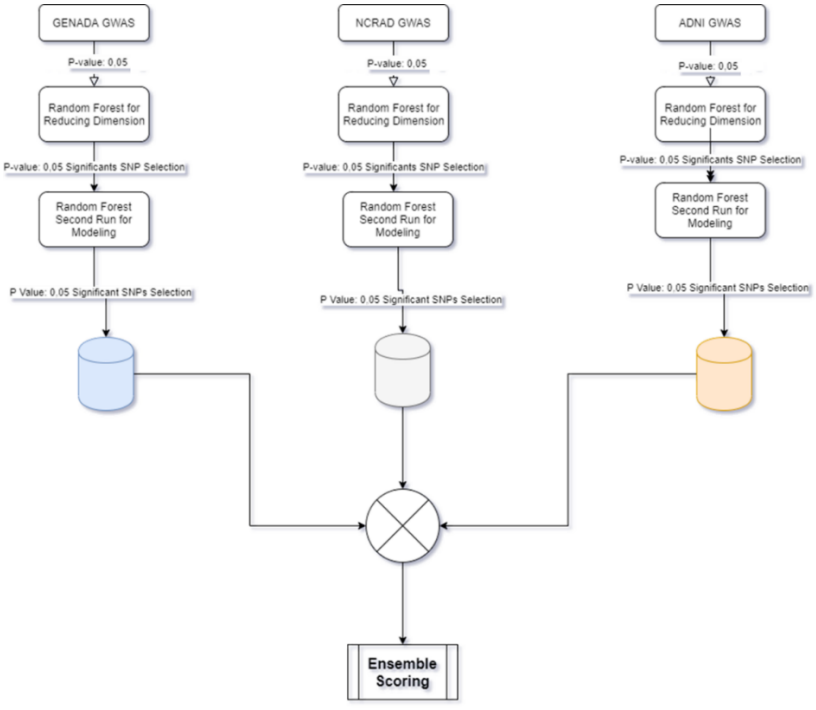


Figure 5: Overview of the Modeling Methodology

4.1. Datasets

The Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu) and NCBI dbGAP (https://www.ncbi.nlm.nih.gov/gap) provide the data used in this study. The NCBI supports and publishes dbGAP (Database of Genotypes and Phenotypes), which was designed to store and exchange genotype and phenotypic data, family history, and other clinical findings of the association studies on a certain disease. Conversely, ADNI (The Alzheimer's Disease Neuroimaging Initiative) is an ongoing, longitudinal, multicenter study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of Alzheimer's disease (AD). To forecast Alzheimer's disease, ADNI gathers, verifies, stores, and uses data from several sources, including genetics, blood biomarkers, Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) scans, and cognitive tests. These datasets provide a comprehensive foundation for examining the genetic underpinnings of LOAD. The integration of data from these varied sources allows for a robust analysis by leveraging large and diverse sample sizes, thereby enhancing the reliability and generalizability of the in-silico models. This approach facilitates a detailed exploration of genetic associations and potential biomarkers for LOAD, contributing to the advancement of our understanding of the disease's etiology and progression.

Multimodels are constructed and assessed using the controlled accessible GWAS datasets from dbGAP projects (1310 controls and 1289 cases), GenADA (777 controls and 798 cases), and ADNI (210 controls and 344 cases). As shown in Table 1, these projects use the Affymetrix Mapping250K_Nsp and Mapping250K_Sty, Illumina Human610_Quadv1_B 500K, and Illumina Human610-Quad BeadChip platforms. The study includes 620.901, 410.907, and 585.295 QC passed SNVs for GenADA, NCRAD, and ADNI, respectively.

Table 1: Databases, Case/Control Groups, and Genotype Platforms

	GenADA	***NCRAD***	***ADNI***
# cases	777	1289	344
# controls	798	1310	210
SNV Platform	Affymetrix Mapping250K_Nsp and Mapping250K_Sty platform	Illumina Human610_Quadv1_B 500K platform	Illumina Human610-Quad BeadChip platform

Datasets must be processed separately due to the differences across platforms and single nucleotide variations focused on AD for determining the causative variants. Thus, before the multi-platform dataset's ensemble process, the individual prediction model finds distinct and noteworthy prior data.

4.2. Evaluation of Random Forest In Terms of Performance Improvement in Pilot GenADA

In our previous research [22] using the GenADA dataset, statistical and significant representative genetic variations' set was obtained by PLINK and Analytical Hierarchy Process (AHP) Scoring method before model construction with a decision tree (ID3, C4.5). Firstly, the same dataset is analyzed by the Random Forest algorithm. The final set which is the result of Random Forest is given as input to ID3 and Random Forest algorithm respectively for predictive disease model construction. Hence, improvements in accuracy via the Random Forest algorithm are tested to be used for the rest of the analysis.

Model accuracies are tested by cross-validation technique and the results are benchmarked with the preliminary study's accuracy rate [22]. If there is an increase, this means Random Forest identifies nonlinear statistical patterns within genomic variables better. As a consequence, RF will be employed for three datasets to choose significant variations as a dimension reduction technique. Alternatively, if there is no significant increase in the accuracy observed, the GWAS study will select statistically significant variations and then AHP scoring will prioritize related genetic variations to AD, finally decision tree will construct the model.

To sum up, after filtering variations from GWAS, the SNVs found to be significantly associated with AD are used as the genotype information for further analysis with various modeling methods including GWAS, Analytical Hierarchy Process (AHP) as dimension reduction, and Decision Tree and Random Forest as modeling algorithm or GWAS and RF as dimension reduction and Decision Tree and Random Forest as modeling algorithm. Hence, for the GenADA dataset, different methods for information gathering are implemented. The disease model that performs best for pilot GenADA is determined as GWAS, Random Forest, and Random Forest multimodel implementation respectively. This is one of the first studies that uses ensemble techniques to integrate multimodels after GWAS and RF analysis.

4.3. Design and Implementation of Analytical Workflow for EnSCAN

Initially, Genome-Wide Association Studies (GWAS) are conducted to identify statistically significant single nucleotide variants (SNVs) by using the PLINK tool for each dataset. PLINK is a widely used tool that employs Chi-Square statistical tests. The chi-square test effectively assesses the relationship between categorical variables, such as genotypes (AA, Aa, aa) and illness status (cases and controls), and is used in GWAS to identify significant SNVs associated with a specific disease. Since GWAS primarily reveals independent statistical significances of genetic variations without addressing their direct relevance to the disease, this information is used to filter out non-informative SNVs that do not exhibit a direct association with the condition under study as illustrated in Figure 6.

Following the preliminary dimensionality reduction to exclude statistically insignificant variations based on GWAS results, the focus shifts to analyzing SNVs that demonstrate a significant association with late-onset Alzheimer's disease (LOAD). Our previous research [22] indicates that the Random Forest-Random Forest (RF-RF) approach has outperformed hybrid models in analyzing the GenADA dataset [49], [50], [51]. Given that the most effective modeling strategy involved GWAS-based filtering followed by a two-step RF process, we have extended this approach to develop the LOAD RF-RF models for each of the three datasets analyzed as illustrated in Figure 6.

This methodological extension aims to leverage the advantages of the RF-RF approach across different datasets, thus enhancing the robustness and generalizability of the model. By integrating GWAS filtering with a two-step RF procedure, the developed LOAD RF-RF models are expected to provide a more accurate and comprehensive analysis of SNVs associated with LOAD, improving the identification of relevant genetic markers and potentially advancing our understanding of the genetic underpinnings of the disease.

In the final phase of this study, we integrate information from three distinct machine learning models for each dataset using a novel ensemble approach as illustrated in Figure 6. For the first time, we propose utilizing a broader spatial context, mapping single nucleotide variants (SNVs) to genomic locations on cytogenetic bands. This method involves ensembling multimodel results to aggregate significant insights obtained from both GWAS and Random Forest (RF) analyses.

Given that SNVs selected by RF differ across datasets due to variations in sequencing platforms, these SNVs are mapped to their respective chromosomal positions. This allows for the analysis of neighboring variants and their potential interactions. We introduce an ensemble scoring algorithm designed to quantify the information related to cytogenetic bands that highlight disease-associated and causative variants, as detailed in Figure 6, and the code is presented in APPENDIX C.

The scoring process involves sliding a three-rowed frame within a cytogenetic band that contains SNVs. The frame size of three rows is chosen to correspond with the three datasets utilized. The algorithm calculates scores based on several factors, including the Linkage Disequilibrium (LD) of each SNV, the incidence of variations in all databases, and validation results from the RF model. These factors are crucial for prioritizing and selecting the most significant variants associated with Alzheimer's Disease (AD). By incorporating these elements, the proposed method aims to enhance the identification of critical genetic variations and provide a more comprehensive understanding of their role in the disease.

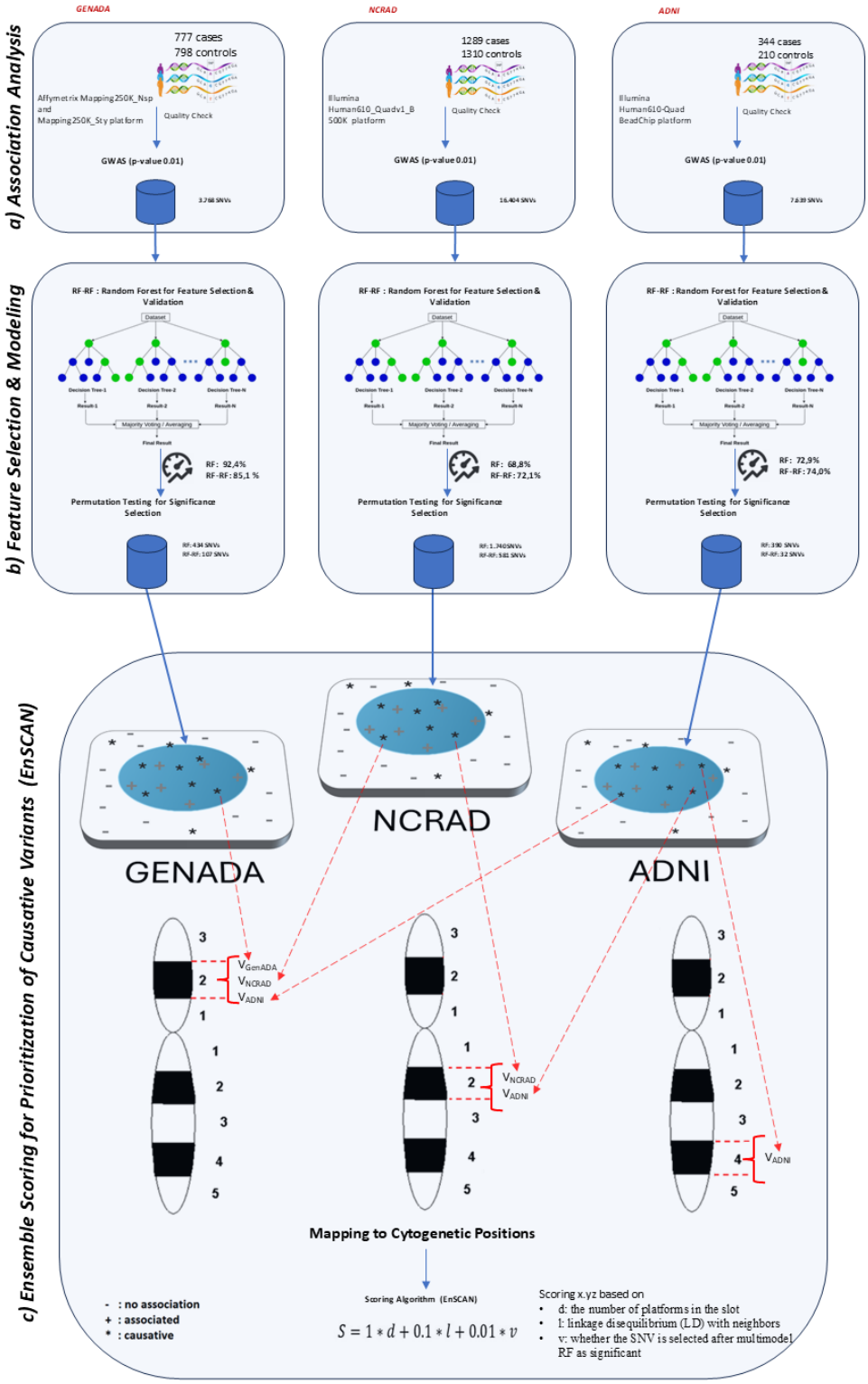


Figure 6: Design for Model Building and EnSCAN

4.3.1. Association Analysis

The data preprocessing step describes any type of operation on data to eliminate noise, clean redundancy and determine the significant subset by feature extraction or dimension reduction methods within the data before modeling.

The single nucleotide variations (SNVs) data provided by the GenADA, NCRAD, and ADNI datasets are examined, and data preprocessing is conducted separately, including data cleaning (identification of missing observations and validation of inconsistent/erroneous data), data integration (combining related data from different sources), data transformation, and data selection (feature selection, dimensionality reduction). Missing values in genetic variables are initially assessed and corrected. Data from distributed sources are consistently integrated.

As allele-calling algorithms continue to improve, quality improvement strategies must ensure that only reliable markers are used for analysis. In data, quality control of genotype data is conducted initially before the authorized access dataset is realized. The quality control and exclusion criteria before the analysis with PLINK for GWAS analysis are performed using the initial parameters such as Hardy Weinberg Equilibrium (poor-quality genotyping can result in heterozygotes being called homozygotes), MAF (low power to detect any association to a disease) and call rate (SNV information is not missing). Autosomal SNVs with Hardy Weinberg Equilibrium value<5e-07, MAF<0.05, and call rate<0.99 are excluded.

To guarantee the accuracy and usability of genotype and phenotypic data, several crucial processes must be completed during the preprocessing stage of data conversion to a relational database. Pulling and importing data from several sources is the first step in the process. Next, genotyping data is systematically transformed and cleaned to remove errors and standardize formats. This entails managing null values, standardizing allele representations, and guaranteeing alignment with the schema of the relational database. Analogously, extensive transformation and cleaning are applied to phenotypic data to rectify inconsistencies, standardize measurement units, and address any difficulties related to data quality. After these procedures, genotype and phenotype datasets are combined to form a coherent dataset that can be analyzed. This process is known as data integration. This thorough preprocessing approach is essential for preserving the integrity of the data and enabling reliable analyses in later studies.

Genome-wide association studies (GWAS) are primarily employed for prefiltering, while the Chi-square test, executed through the PLINK tool, is applied to each dataset independently to obtain p-values for determining the significance of associated single nucleotide variants (SNVs).

Adjusted p-values, such as those modified by Bonferroni or False Discovery Rate (FDR) corrections, are employed in GWAS to account for multiple comparisons. These adjustments are crucial in hypothesis testing and in controlling for false

positives, although they may lead to the exclusion of many potentially informative features. When the FDR threshold for multiple testing is set at a GWAS significance value of 0.0001 (10^{-5}), the number of significant SNVs decreases, resulting in an insufficient input size for machine learning algorithms as seen in Figure 7.

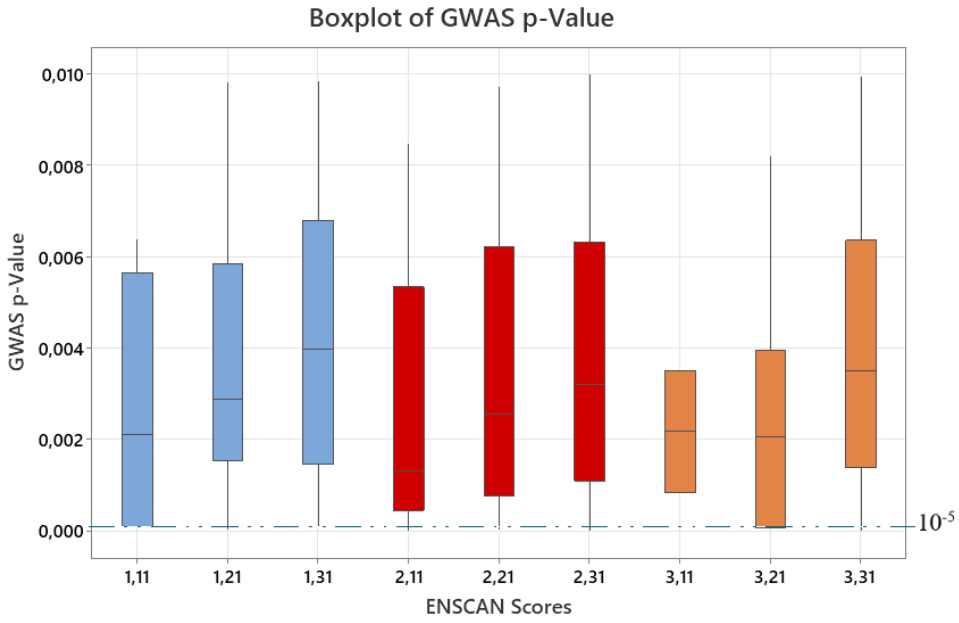


Figure 7: Comparison of GWAS p-values Relative to EnSCAN Scores

Finding predictive features becomes the new objective when modeling with ML algorithms such as RF, instead of finding statistically significant relationships. Adjusted p-values prioritize statistical significance over model optimization, potentially discarding SNVs that could be useful for identifying complex patterns. To mitigate the loss of valuable information while accommodating nonlinear relationships between disease-related SNVs, a Type I error threshold (alpha = 0.01) is utilized. This threshold facilitates the identification of nonlinear associations between SNVs and the disease through machine learning algorithms, enabling the evaluation of these associations' contributions to the model without the need for multiple hypothesis correction.

Data Migration to Relational Database and Pulling & Importing

Data is generally distributed, and it can be obtained from different resources. Before the modeling, data must be integrated. This may require schema integration which also means combining metadata from different data resources. While integrating different databases, redundancy problems both in data or attribute names may arise and it must be resolved. Each study dataset contains genotype data for each case and controls as well as their phenotype data. Individual files are independent and include the genotype data of each subject.

Individual files and their phenotype files are stored as text files which are in the form of human-readable. Each file contains the related sample's id as a comment line in it. Such big, distributed data brings some problems about how to keep and query data. Database management technologies have column limits but there is no such limit for row entries. SQL Server 2016 Enterprise edition is used to migrate all study data. Individual genotype data files and phenotype data files are decrypted and then genotype data is migrated from the file system to database row by row for each individual by matching their related sample's id. Phenotype data is inserted in the database with their related sample ID, as well. Data migration is automatized by a Java program that visits individual files and imports by reading line by line. Steps are given in Algorithm 1 and Algorithm 2 below.

Algorithm 1: Genotype Data Migration from File System to Database

Step 1: Create a table in the database for genotype data.
Step 2: Get the list of each data file name.
Step 3: Read each data file by file.
Step 4: Read individual file line by line.
Step 5: Split each line with a space delimiter.
Step 6: Execute insert scripts for each line.

Algorithm 2: Phenotype Data Migration from File System to Database

Step 1: Get the list of each data file name.
Step 2: Read each phenotype data file by file.
Step 3: Assign a table name in the database for each phenotype data file.
Step 4: Read the phenotype data file line by line.
Step 5: Split each line with a space delimiter.
Step 6: Execute insert scripts for each line.

After creating the database, both SNV data and phenotype data are pulled into memory by using R for further investigation of data. ODBC connection is created via "RODBC" package and data is imported to the R environment by having query language (SQL).

Data Handling (Transformation and Cleaning) For Genotype Data

SNV genotype data is stored row by row for each individual and must be transformed to the shape for further analysis. To process data by random forest algorithm, the individual's sample's ID must be transformed into rows as well and SNV genotype

attributes must be transformed into rows. At the end of the transformation step, $(m\ x\ n)$ data frame will be obtained where n is the number of samples and m is the number of SNVs.

Since SNV genotype data is obtained from SNV chips, this transformation process is only possible in memory and unable to be written to a file system. R program has a library to handle the big data transformation process. The package is called "reshape2".

Reshape package transposes the data according to a given condition using the "dcast" function. In the database, reshaping is implemented as SNVs are columns and samples are rows.

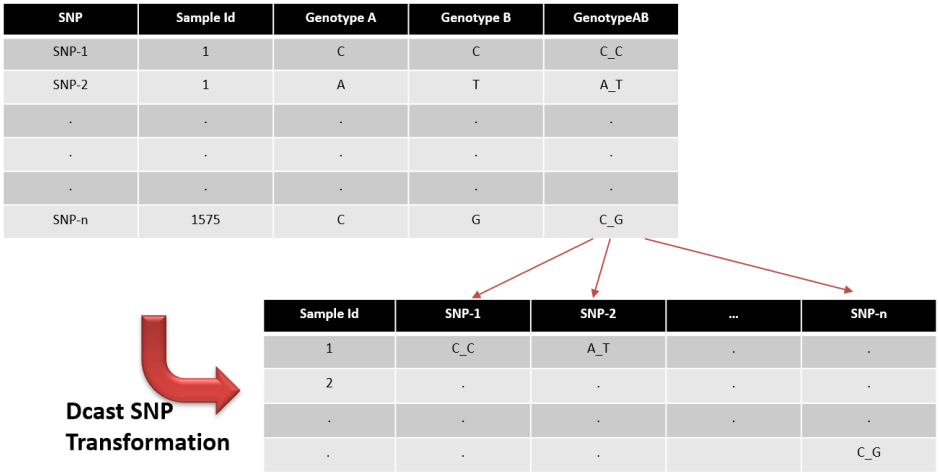


Figure 8: Reshape2 package: Dcast function Implementation

SNV alleles are formed of AB based on a pair of alleles carried by individuals. If two alleles are in the same state such as AA, CC, GG, and TT, these are called homozygotes. If the two alleles are different such as AC, AG, AT, CG, CT, and GT, these are called heterozygotes. Heterozygotes must give the same information if there is no maternal or paternal discrimination between, for example, AC and CA. Hence, one of the important transformations is recording SNVs.

SNV attributes whose values are null after initial data and dimension filtering which means SNV or sample call rates are below the threshold, are checked for filling in. As SNVs are categorical data, most frequently SNV allele value which is also the mode of the distribution is replaced by missing values.

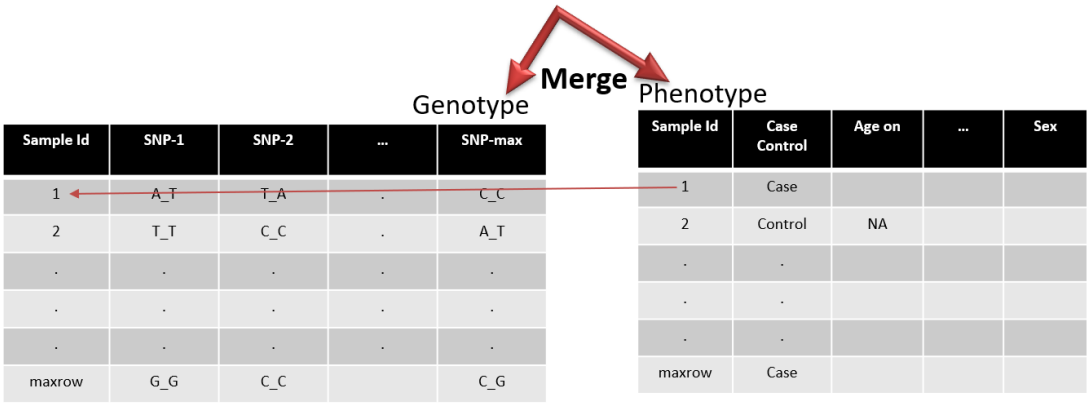
Data Handling (Transformation and Cleaning) For Phenotype Data

Firstly, phenotype data is controlled if it contains null values. After null control, transformation is implemented to phenotype phenotype-dependent variable which is the status of the disease. Case and control groups make up the dependent variable's categories. In datasets, they are given as numbers which number 1 stands for cases and 0 stands for controls.

Data Integration

Genotype data and phenotype attributes are kept in different tables in SQL Server, tables are merged to obtain required data for further analysis by using the sample IDs Identical sample IDs can be taken into account for the merge operation of different tables.

To get the analyzed data, transformed data is used for integration. In the R program, the merge function is used to join different tables.



Sample Id	SNP-1	SNP-2	...	SNP-max
1	A_T	T_A	.	C_C
2	T_T	C_C	.	A_T
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
maxrow	G_G	C_C		C_G

Sample Id	Case Control	Age on	...	Sex
1	Case			
2	Control	NA		
.	.			
.	.			
.	.			
maxrow	Case			

Figure 9: Merge SNV data and Phenotype for Each Dataset

To develop a model with high accuracy and meaningful results, it is essential to first preprocess and prepare the data for further analysis.

Elimination of Redundant Variations

GWAS analysis is implemented for each dataset to eliminate redundant variations by comparing the results between unadjusted p-values with a significant value of 0.01. GWAS is used only for pre-filtering; hence, multiple hypotheses are not considered.

4.3.2. Feature Selection and Model Building: Individual Dataset Dimensionality Reduction and Modeling by Random Forest

Dimensionality reduction can enhance both computational efficiency and the accuracy of machine learning models. A robust method for achieving dimensionality reduction involves constructing an extensive set of decision trees targeting a specific attribute and then utilizing the usage statistics of each attribute, referred to as variable importance, to identify the most informative subset of features using the Random Forest technique. One advantage of this approach is that it does not require pre-assumptions or preprocessing steps, such as outlier detection. The Random Forest algorithm mitigates overfitting by constructing a large ensemble of de-correlated trees and aggregating their predictions through voting.

Random Forest is a widely used method for feature ranking. An attribute that frequently serves as the optimal split for decision trees in the ensemble is likely to be an informative and significant feature. Variable importance scores, calculated within the Random Forest, indicate which attributes are the most predictive for the problem at hand. These scores are derived from two measures: Gini gain, which is the average improvement in impurity produced by a variable across all trees, and Permutation Importance, which reflects the mean decrease in classification accuracy resulting from permuting the variable, averaged over all trees. Random Forest ranks variables based on permutation importance scores in descending order to select the most significant variables.

To discover the best performance that Random Forest generates, parameters should be tuned instead of using default ones. It is also crucial for enhancing predictive performance and identifying important variables. However, it is a difficult process to find a good or even well-performing algorithm parameter through a process of trial and error. Instead, a different combination of parameters builds up different forests and calculates the accuracy for each. In this way, what effects the parameters have on the model accuracy is known. Based on accuracy that is derived from the recorded OOB error rate, trees are assessed where the out-of-bag error rate stabilizes and reaches the minimum.

Three primary features should be adjusted to improve model accuracy.

1. ***Mtry* (maximum features):** This is the number of randomly selected variables that Random Forest is allowed to try in each tree. This value is generally the square root of a number of variables. It can be estimated by log 2 as well. In addition to these calculations, it is also possible to be defined as ½ times of square root of a number of variables or 2 times of square root of variables. Increasing the *mtry* value seems to improve accuracy but decreases the diversity (amount of randomness between trees) of individual trees in the forest. Hence, optimal *mtry* with the right balance should be obtained.
2. ***Ntree* (*n* estimators):** This is the number of trees before taking the majority voting. Higher trees may reach better performance in terms of accuracy, but it affects Random Forest run-time.
3. **Leaf Size (depth):** Leaves are the end nodes of a decision tree. A smaller leaf makes the model more prone to capturing noise in data. Random Forest, does not require pruning of the trees. Hence, minimum leaf size must be preferred.

Following the Genome-Wide Association Study (GWAS) conducted using PLINK, the Random Forest (RF) algorithm is applied in conjunction with dimensionality reduction and the implementation of a supervised learning model [46].

The dataset is partitioned using bagging (bootstrap aggregating) techniques to minimize prediction variance by aggregating the results from multiple classifiers

trained on different sub-samples of the original dataset. This involves sampling from the original data to create new datasets for constructing each decision tree. Data not included in the training samples is termed "out-of-bag" (OOB) and is used to evaluate the accuracy of the individual trees.

Initially, parameter tuning for the Random Forest model involves determining the number of trees to generate and the number of features to randomly consider for each split. Different values for the "number of trees" are tested to identify the optimal size, while the "number of randomly selected variables" is also adjusted. The "RANGER" package in R, which includes functionality from the "randomForest" package, is utilized for this tuning process. The RANGER library automates the fitting of multiple versions of an ensemble model by varying key parameters and tree construction samples, thus facilitating the optimization of Random Forest parameters. Although increasing the "number of trees" enhances computational effort, it often leads to improved model accuracy. Sequential values for these parameters are tested, and the model is trained accordingly.

A confusion matrix is then constructed to evaluate general accuracy, specificity, and sensitivity, based on majority voting from the OOB samples. The model with parameters that yield the highest accuracy in terms of OOB performance is selected as the final model.

Finally, the importance of each variable is assessed using the RANGER package in R by calculating t-test statistics for each variable in the most accurate model. This process determines the contribution of each variable to the model's predictive power, providing insight into their relative importance within the model.

4.3.3. Ensemble Scoring (EnSCAN) for Prioritizing Causative Variants

In the literature, several ensemble methods have been proposed for use within the same dataset [14], [29], [16], [52], [53]. However, to date, there has been a lack of ensemble methodologies designed for analyzing data across multi-platforms with differing attributes. This study introduces a novel approach wherein the outputs from initial prediction models are employed as prior knowledge to integrate and synthesize information from various sources, thereby computing an importance score for each single nucleotide variant (SNV) to assess its relevance to Late-Onset Alzheimer’s Disease susceptibility.

As the innovation and novelty of the Ensemble Scoring Algorithm, we have utilized genomic locations where all RF-selected SNVs are mapped to their chromosomal positions in the genomic space, and cytogenetic bands are set as slots to survey through a sliding window approach for scoring as seen in Figure 10. In this study, a "window" is defined as a particular segment of the genomic sequence under analysis. This segment is characterized by a specific number of base pairs, within which variations of interest can be identified and examined. As tagged SNVs do not overlap across different

platforms, we developed a framework that maps each SNV to its chromosomal location on genes, allowing for the observation of neighboring variants across various datasets.

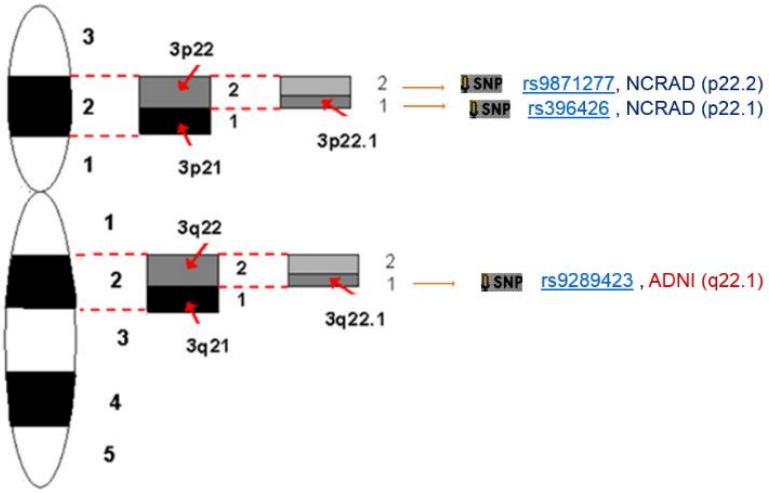


Figure 10: Mapping to Cytogenetic Locations (Genomic Positions) on Chromosomes

This methodological advancement allows for a comprehensive evaluation of SNVs by incorporating spatial genomic information, thus providing deeper insights into their potential role in LOAD. By leveraging the spatial context of genomic locations and employing a systematic scoring mechanism, this approach aims to enhance the accuracy and interpretability of genetic associations with the disease.

The Ensemble Scoring Algorithm takes into account three factors: (1) the number of platforms in the scoring window that demonstrate variants from various data sets are concentrated in the same region; (2) linkage disequilibrium (LD) with neighbors; and (3) whether the SNV is chosen as significant following multi-step RF Based on these three variables, each variant is assessed by exploiting the neighboring SNVs in the sliding window, which is a method to define SNVs' scores. A higher score for an SNV first signifies that numerous unlinked SNVs are connected with that locus and that the loci are linked to the condition in various studies. Using a multi-step RF-RF algorithm, an additional SNV selection identifies the potential causal variations inside that window.

The "sliding window" is a tool employed to scan the SNV list, which is organized according to cytogenetic locations. The goal is to evaluate each SNV based on various factors. One of these evaluation factors is the database listing, with its maximum value corresponding to the total number of databases utilized in the study.

The number of SNVs in each sliding window is set to three to analyze multiple datasets. Figure 11 represents the Ensemble Scoring Algorithm's flowchart, which is used to calculate the importance of SNVs.

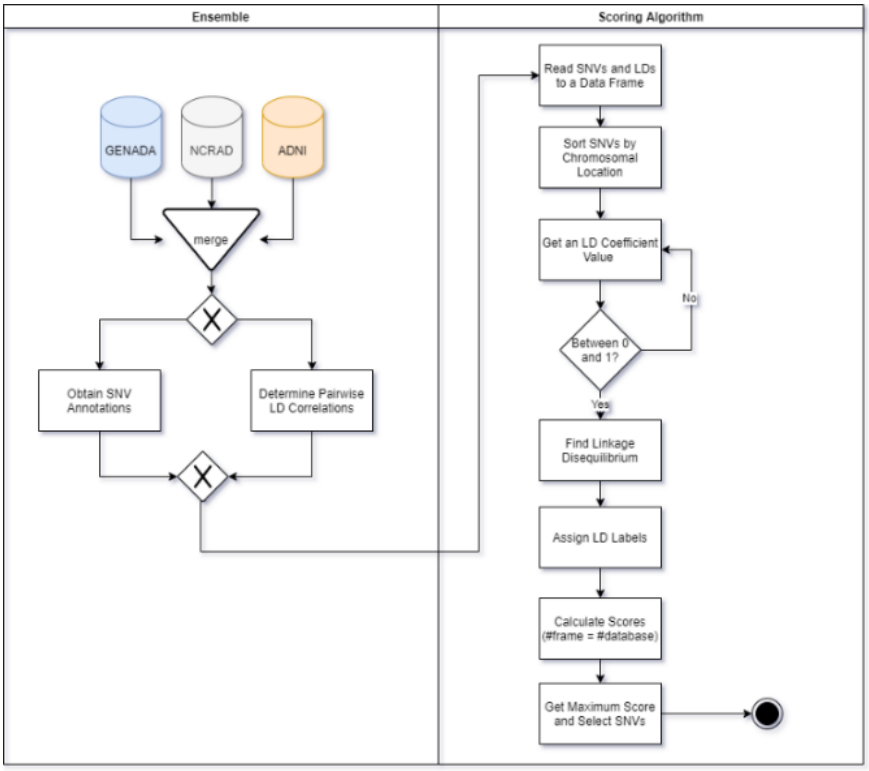


Figure 11: Flowchart of Ensemble Scoring Algorithm (EnSCAN)

The window size is determined by the number of datasets used in the study and its size is equal to the number of datasets. During the analysis, for some SNVs, the window can be extended if there is a linkage between the target SNV and its successor SNV in the list. To analyze the entire list of SNVs, the sliding window is moved by one unit starting from the top of the list. So, the step size is 1. In this way, each SNV can be listed in several windows at different locations at each time and each SNV can be evaluated multiple times by using its neighboring information in each window.

The steps of the algorithm are given in Algorithm 3. The Pseudo Algorithm is also detailed and given in APPENDIX A. All SNVs selected by the RF model are scored using the Ensemble Scoring Algorithm. In Step 6, each SNV is scored three times, as listed in different windows.

Algorithm 3: Ensemble Scoring Algorithm

Step 1: All SNVs from different databases selected by the RF models were merged.

Step 2: Sort SNVs according to chromosomal location.

Step 3: Set a minimum LD coefficient and mark whether an LD linkage exists for neighboring SNVs.

Step 4: The SNVs selected by RF-RF modeling are marked.

Step 5: Browse the SNV list using a sliding window scheme starting from the first SNV.

Step 6: Score each SNV covered by the window as in Formula 1. Each SNV will have three scores listed in three consecutive windows.

Step 7: Assign the maximum of three score values as the importance score of each SNV.

The scoring formula for each SNV has three components: one from database listing, one from LD linkage, and the final component refers to whether the SNV is selected by the RF-RF model. The Ensemble Score of an SNV is calculated with the formula below:

$$Score\ (SNV) = (1 * d) + (0.1 * l) + (0.01 * v) \qquad \text{(Formula 1)}$$

The score is a representation for classifying an SNV using an evaluation based on three aspects: (i) database listing, (ii) LD linkage, and (iii) RF-RF modeling result. Here, there is a hierarchy between those three aspects of evaluation. As the main goal of the study is to ensemble results from independent GWAS studies, database listing has a higher hierarchy (can also be interpreted as importance) than the other two aspects, LD linkage and RF-RF modeling results have an equal hierarchy where LD linkage has priority in evaluation. To represent this evaluation structure, a three-digit system of representation is used. To present the hierarchy the first digit (which denotes the score of the database listing) is separated by (.) notation. Ordering information between LD linkage and RF-RF modeling result is represented by the second and third digits where the second digit denotes precedence in evaluation.

Given that the analysis involves three datasets, the variable *d* can assume values of 1, 2, or 3. The parameter *l* is determined based on the number of independent single nucleotide variants (SNVs) present within the defined sliding window and can therefore be 1, 2, or 3. If both pre-existing and post-existing neighbors are in Linkage Disequilibrium (LD), the value of *l* is assigned the minimum score of 1. The variable *v* indicates whether an SNV has been selected through the Random Forest-Random Forest (RF-RF) approach. Specifically, *v* is set to 1 if the SNV is identified as significant by RF-RF.

When evaluating single nucleotide variants (SNVs) across three distinct multi-platform datasets, the scoring system yields a maximum possible score of 3.31 and a minimum score of 1.10 for any given SNV. This scoring range reflects the extent to which each SNV is associated with the disease of interest, integrating information from multiple datasets and the results of various analytical models. A higher score for an SNV first indicates that the locus is associated with the condition according to multiple studies, and multiple unlinked SNVs are associated with that locus. Further selection of an SNV

through a multistep RF-RF model revealed the candidate causative variants in that window.

The maximum score of 3.31 represents the highest level of significance or relevance an SNV can achieve, indicating strong support from all datasets and robust associations with the disease. Conversely, a minimum score of 1.10 denotes the lowest level of significance, suggesting weaker or less consistent evidence across the datasets. In summary, the score can be an element of S, where S ∈ {1.10, 1.11, 1.20, 1.21, 1.30, 1.31, 2.10, 2.11, 2.20, 2.21, 2.30, 2.31, 3.10, 3.11, 3.20, 3.21, 3.30, 3.31}.

All SNVs are scored using the Ensemble Scoring Algorithm. Each SNV is scored three times as it is listed in different windows.

“rs1333190” is considered as an example. The scoring procedure is visualized as shown below for each location of the SNV in the windows.

CASE 1: SNV is in the middle of a sliding window

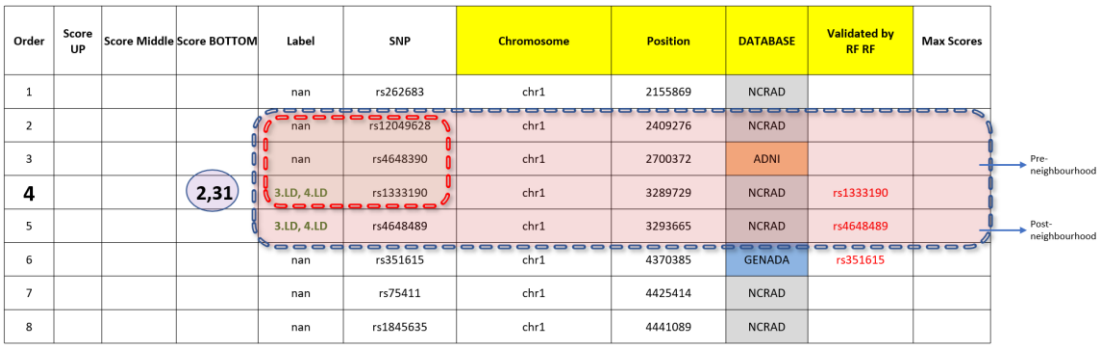
Order	Score UP	Score Middle	Score BOTTOM	Label	SNP	Chromosome	Position	DATABASE	Validated by RF RF	Max Scores
1				nan	rs262683	chr1	2155869	NCRAD		
2				nan	rs12049628	chr1	2409276	NCRAD		
3				nan	rs4648390	chr1	2700372	ADNI		
4		3,21		3.LD, 4.LD	rs1333190	chr1	3289729	NCRAD	rs1333190	
5				3.LD, 4.LD	rs4648489	chr1	3293665	NCRAD	rs4648489	
6				nan	rs351615	chr1	4370385	GENADA	rs351615	
7				nan	rs75411	chr1	4425414	NCRAD		
8				nan	rs1845635	chr1	4441089	NCRAD		

Figure 12: SNV is located in the middle of the window.

Figure 12 illustrates the scoring calculations for the single nucleotide variant (SNV) identified as "rs1333190," which is positioned centrally within a three-slot sliding window. According to Step 6 of Algorithm 3, the sliding window is constructed with "rs1333190" at its center. The window is extended due to the presence of Linkage Disequilibrium (LD) associated with "rs1333190." Consequently, information from three distinct databases is incorporated (ADNI, NCRAD, GenADA) within the expanded window.

The term "tandem" refers to the integration of data from these three separate databases. Thus, the parameter *d* is assigned a value of 3, reflecting the utilization of information from all three datasets. Given the presence of LD linkage for "rs1333190," the parameter *l* is set to 2 within the frame. Additionally, because "rs1333190" was selected by the Random Forest-Random Forest (RF-RF) algorithm, the parameter *v* is assigned a value of 1. Consequently, the final score for "rs1333190" is calculated as 3.21, as detailed in Formula 1.

CASE 2: SNV is at the bottom of a sliding window



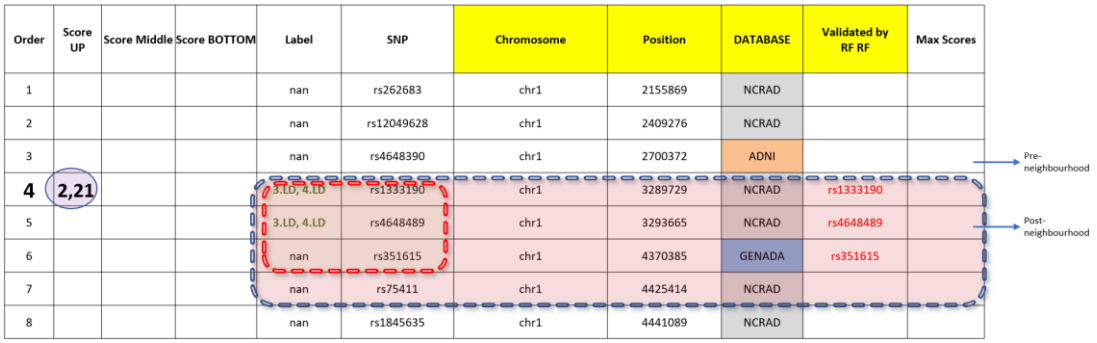
Order	Score UP	Score Middle	Score BOTTOM	Label	SNP	Chromosome	Position	DATABASE	Validated by RF RF	Max Scores
1				nan	rs262683	chr1	2155869	NCRAD		
2				nan	rs12049628	chr1	2409276	NCRAD		
3				nan	rs4648390	chr1	2700372	ADNI		
4			2,31	3.LD, 4.LD	rs1333190	chr1	3289729	NCRAD	rs1333190	
5				3.LD, 4.LD	rs4648489	chr1	3293665	NCRAD	rs4648489	
6				nan	rs351615	chr1	4370385	GENADA	rs351615	
7				nan	rs75411	chr1	4425414	NCRAD		
8				nan	rs1845635	chr1	4441089	NCRAD		

Figure 13: SNV is located at the bottom of the window.

In the first step, as described in Step 6 of Algorithm 3, the sliding window is positioned with "rs1333190" at the bottom. Figure 13 illustrates the scoring calculation when the SNV is situated at the bottom of a three-slot window. In this scenario, "rs1333190" is placed at the bottom of the window, and due to its linkage with "rs4548489," the sliding window expands to include both as one informative variant.

Here, the parameter *d* is assigned a value of 2, reflecting the inclusion of data from two distinct datasets (NCRAD, ADNI) within the expanded window. Given that there is no Linkage Disequilibrium (LD) linkage present within the frame for the additional variant, the parameter *l* is set to 3. Additionally, since "rs1333190" is selected by the Random Forest-Random Forest (RF-RF) algorithm, the parameter *v* is given a value of 1. Thus, the final ensemble score for "rs1333190" is calculated as 2.31, as detailed in Formula 1.

CASE 3: SNV is at the top in the sliding window



Order	Score UP	Score Middle	Score BOTTOM	Label	SNP	Chromosome	Position	DATABASE	Validated by RF RF	Max Scores
1				nan	rs262683	chr1	2155869	NCRAD		
2				nan	rs12049628	chr1	2409276	NCRAD		
3				nan	rs4648390	chr1	2700372	ADNI		
4	2,21			3.LD, 4.LD	rs1333190	chr1	3289729	NCRAD	rs1333190	
5				3.LD, 4.LD	rs4648489	chr1	3293665	NCRAD	rs4648489	
6				nan	rs351615	chr1	4370385	GENADA	rs351615	
7				nan	rs75411	chr1	4425414	NCRAD		
8				nan	rs1845635	chr1	4441089	NCRAD		

Figure 14: SNV is located at the top of the window

Figure 14 presents the scoring calculations for the single nucleotide variant (SNV) designated "rs1333190" when positioned at the top of a three-slot sliding window. Following Step 6 of Algorithm 3, the sliding window is arranged such that "rs1333190" remains at the top. The window is extended due to the presence of Linkage Disequilibrium (LD) associated with "rs1333190."

In this context, the parameter *d* is assigned a value of 2, indicating that data from two distinct datasets (NCRAD, GenADA) are included within the expanded window. Since there is an LD linkage for "rs1333190" within the window, the parameter *l* is set to 2. Additionally, because "rs1333190" is selected by the Random Forest-Random Forest (RF-RF) algorithm, the parameter *v* is assigned a value of 1. Consequently, the final score for "rs1333190" is calculated as 2.21, as defined in Formula 1.

The single nucleotide variant (SNV) "rs1333190" is evaluated across all three possible positions within the sliding window. It receives scores of 3.21 when located at the center of the window, 2.31 when positioned at the bottom, and 2.21 when placed at the top. In this instance, the ensemble score for "rs1333190" is determined to be 3.21, which is the highest score among the three calculated EnSCAN scores. This scoring technique has been applied to all single nucleotide variants (SNVs), and calculations have been conducted accordingly.

4.3.4. Variance Annotation Process

The assignment of functional information to genetic variants is known as variant annotation. This process includes evaluating sequence conservation, predicting the impact of variants on protein structure and function, and genomic mapping. In our analysis, we utilized the SNPnexus and SNiPA tools to annotate the filtered variants. SNPnexus [54] was utilized to annotate the single nucleotide variants (SNVs). This web-based tool is designed for the annotation of genetic variants, assisting in the selection and prioritization of both known and novel genomic alterations. We annotated our SNV lists using SNPnexus to query the Genetic Association Studies (GAD) and the NHGRI Catalogue of Published Genome-Wide Association Studies (GWAS) databases, enabling comparison of associated genes and diseases. The human assembly GRCh37/hg19 was selected, with the European (non-Finnish) minor allele frequency (MAF) used as the population reference. Additionally, SNiPA was employed to assess pairwise linkage disequilibrium (LD). The correlation between the phenotype and the marker allele suggests the potential presence of the causal SNV in LD. By encompassing all LD blocks within the genomic region, additional information was obtained to identify variants that may predispose individuals to the disease. The LD correlation coefficient was determined during the scoring process, and the LDs of each SNV were defined and labeled accordingly.

Finally, all SNVs identified in the LOAD RF-RF model, using three platforms with multistep models, were prioritized through the sliding window algorithm for scoring.

4.3.5. Functional Enrichment

After the GWAS algorithm is run for the ensemble and scores for individual SNVs are calculated, the pathway enrichment and visualization of the data are done by following protocol [55]. Pathway enrichment analyses are done by g: Profiler, a web server accessible to the public for characterizing and manipulating lists of genes. Functional

enrichment analysis (over-representation analysis (ORA) or gene set enrichment analysis on AD-related SNVs) is performed to map genes to known functional information sources and detect statistically significantly enriched terms.

The g: GOSt component of the g: Profiler tool is used to analyze variants found by the ensemble model, with a particular emphasis on the Reactome pathways, GO Molecular Function, GO Cellular Component, and GO Biological Process. All analyses employ default attributions and pathways linked to SNVs based on ensemble scores are identified using a significance level of 0.05. Fisher's exact test is used to calculate the enrichment p-values for pathways, and the Bonferroni correction procedure is used to apply multiple testing corrections.

Next, functional enrichment visualization is achieved using EnrichmentMap, a Cytoscape plugin, which builds networks from Reactome pathways and Gene Ontology annotations. The p-value threshold of 0.05, FDR q-value cut-off of 0.01, and edge similarity cut-off of 03 (Jaccard metric) are used in all analyses.

CHAPTER 5

RESULTS

This study presents an approach for prioritizing SNVs based on prior knowledge obtained from different datasets for the early diagnosis of Alzheimer's disease through the genome-scale ensemble method.

A machine learning classification study for the differential diagnosis of Alzheimer's disease was conducted by using a single dataset GenADA in our previous study [22]. In this previous study, PLINK and AHP scoring by METU SNP was employed for dimension reduction, and the CART algorithm was run to detect significant SNVs in terms of biological and statistical significance after PLINK analysis. Alternatively, CART is employed soon after PLINK analysis considering statistical significance.

Considering the performance improvement in accuracy, a multimodel (RF-RF) approach is employed to detect the non-linear relationships of disease-associated genes and for dimensionality reduction. Since the accuracy rate of the multimodel shows an increase compared to the accuracy rate of the previous study [22], dimensionality reduction and individual dataset modeling are completed using multi Random Forest classification algorithm to identify significant variation for each dataset (NCRAD and ADNI). Finally, the Ensemble Scoring Algorithm (EnSCAN) is prioritized by SNVs based on prior information captured from distinct datasets. Results are given in the subheadings.

5.1. Accuracy Results of Random Forest Algorithm with Different Train Parameters on GenADA

In our previous study [22], we first selected statistically significant SNVs using PLINK by assessing their correlations. To streamline the data, we applied the Analytic Hierarchy Process (AHP) for dimensionality reduction. Hence, we prioritized these SNVs based on p-values in terms of biological meaning with METU SNP, which helped in selecting a subset of representative SNVs. We then created a decision tree for the predictive diagnose model. Our results showed minimal difference in diagnostic accuracy between the two approaches: the SNV-only decision tree had an accuracy of 56.08% by PLINK – AHP – ID3.

For the GenADA dataset as a pilot after PLINK, Random Forest is implemented to compare the accuracy with the ID3 model after AHP selects SNVs. For dimension reduction and modeling purposes, RF is employed, and results are explained in a detailed way in Figure 15 and Figure 16. The result shows that the accuracy of the first-run Random Forest after PLINK is 92.44%. Considering the statistically significant variables via permutation test (RANGER function in R), second-run Random Forest as a validation modeling is employed. At the end of the second-run Random Forest over GenADA dataset, the model accuracy decreased to 85.08% as expected. This value is still noteworthy compared to modeling after AHP by PLINK – RF – RF.

After determining the statistically significant SNVs by PLINK, Random Forest multimodel implementation for all datasets is decided to be implemented due to observed improvements in performance in the pilot dataset (GenADA) instead of using AHP scoring and decision tree modeling. Hence, we aim to investigate whether Random Forest is a better way of selecting significant variables to lay suitable groundwork before the ensemble.

5.2. Results for GWAS Associated with LOAD

We constructed in-silico models for Late-Onset Alzheimer's Disease (LOAD) utilizing genotyping data derived from three distinct datasets obtained from the ADNI and dbGAP initiatives. Our analysis encompasses Genome-Wide Association Study (GWAS) data from the following sources: ADNI, which includes 210 control subjects and 344 cases; GenADA, comprising 777 control subjects and 798 cases; and NCRAD, accessed through dbGAP, with 1310 controls and 1289 cases.

In the first step, GenADA, NCRAD, and ADNI datasets are analyzed independently, and after preprocessing, PLINK is used for GWAS followed by p-value filtering for the initial dimension reduction.

The initial phase of the analysis involves performing Genome-Wide Association Studies (GWAS) using the PLINK software. This is followed by dimension reduction through filtering based on p-values to identify statistically significant single nucleotide variants (SNVs). For the analysis of individual SNVs, a type I error rate of $\alpha = 0.01$ is applied.

Post-GWAS, a total of 3.768 SNVs from the GenADA dataset, 16.404 SNVs from the NCRAD dataset, and 7.639 SNVs from the ADNI dataset are identified as statistically significant for Late-Onset Alzheimer's Disease (LOAD) [46]. To refine these findings, PLINK is employed to filter out redundant SNVs that are either common across different cohorts or deemed uninformative. This filtering process ensures that unique and relevant SNVs are retained for further analysis, thereby improving the specificity and accuracy of the results in identifying genetic markers associated with LOAD.

For each dataset, a Multi-step Random Forest (RF) is executed with 5-fold cross-validation (CV) using the RANGER R package, following GWAS with PLINK.

5.3. Model Construction and Tuning for Multi Platforms by Random Forest

Following the GWAS, dimensionality reduction is performed using the Random Forest algorithm across three datasets: GenADA, NCRAD, and ADNI. The optimization of the *mtry* and *ntree* parameters is carried out for each dataset to identify the best model configurations based on out-of-bag (OOB) sample accuracy.

The result from the Random Forest analysis, utilizing permutation-based significance, is used as input variables for a second round of Random Forest modeling. This approach is employed to validate these parameters and assess whether they enhance the accuracy of the model by tuning.

The graph shows the number of trees (*ntree*) on the x-axis and model accuracy on the y-axis. Four lines are drawn, one for each randomly selected variable (*mtry*) value. The graph shows the general trends in the increase in performance with *ntree* and that the larger value of *mtry* is probably preferred.

For the GenADA RF-RF approach, dimensionality reduction on 3,768 significant SNVs results in a final model with parameters *mtry*=122 and *ntree*=1,000, achieving an accuracy rate of 92.44% as Figure 15 illustrates.

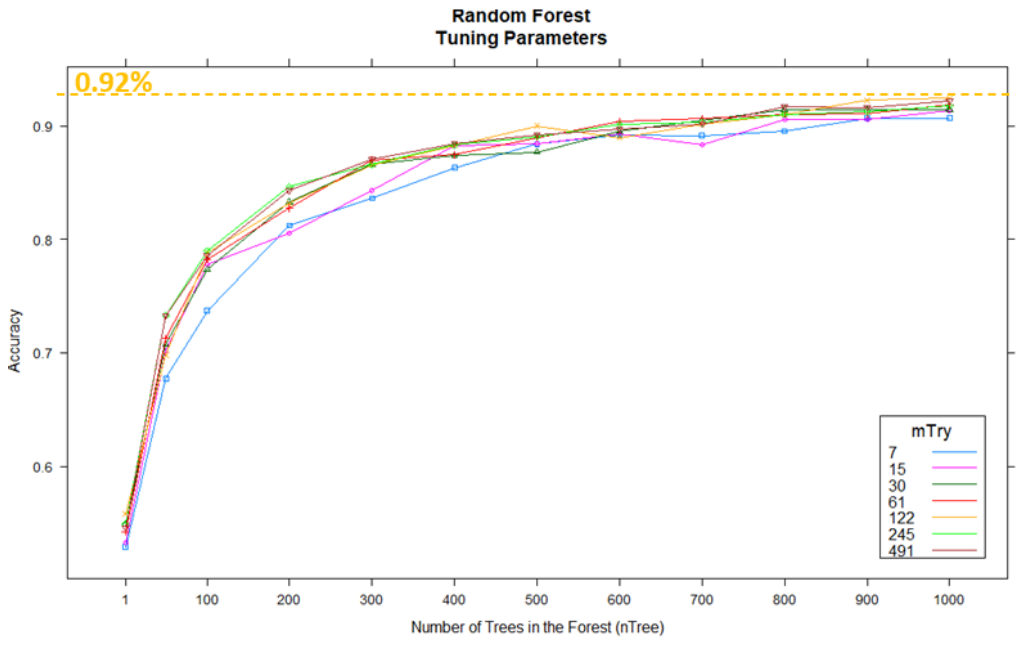


Figure 15: GenADA RF Tuning Parameters

Permuted significance testing was conducted with 100 permutations using the RANGER package. Permutation testing identifies 434 statistically significant SNVs, which are used in a second Random Forest model (GenADA RF-RF), yielding a tuned model with *mtry*=2 and *ntree*=900, achieving 85.08% accuracy as shown in Figure 16 and identifying 107 significant SNVs.

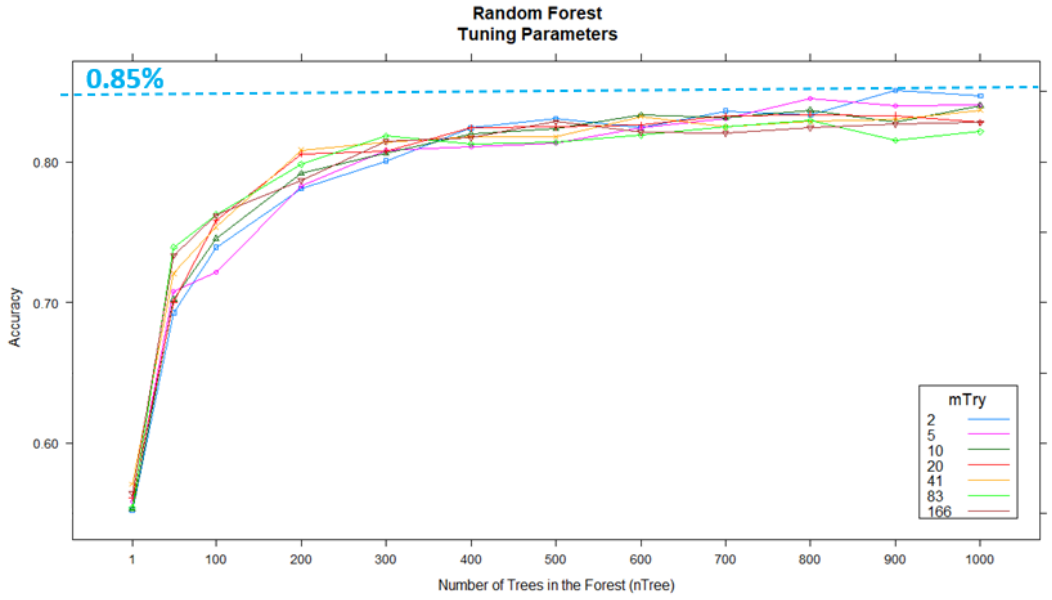


Figure 16: GenADA RF-RF Tuning Parameters

For the NCRAD RF-RF approach, the analysis starts with 16,404 significant SNVs. The best model, after tuning, has parameters *mtry*=1,024 and *ntree*=1,200, resulting in 68.8% accuracy as Figure 17 visualizes.

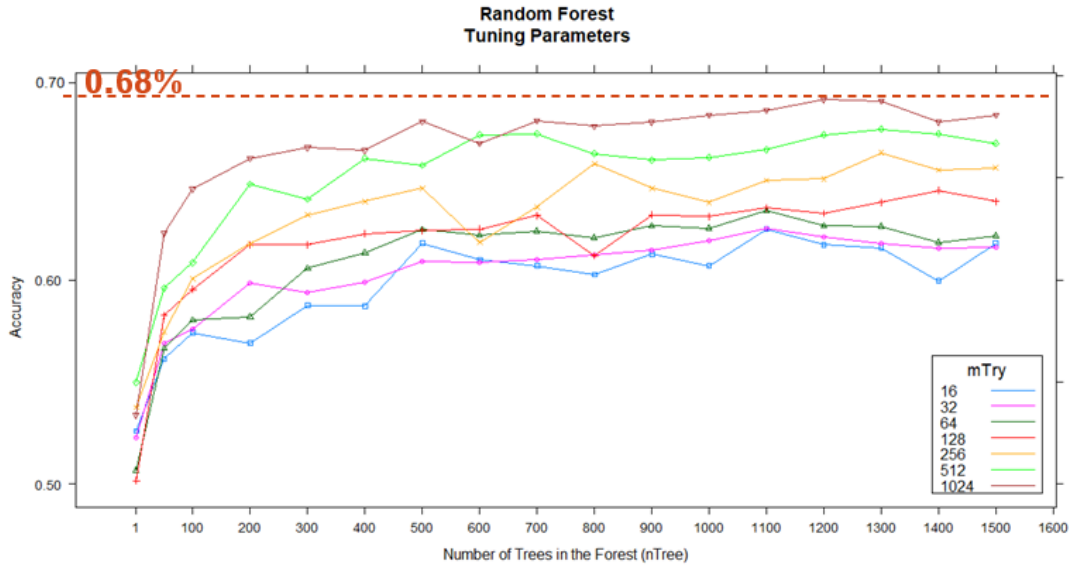


Figure 17: NCRAD RF Tuning Parameters

Permuted significance testing was conducted with 100 permutations using the RANGER package. Permutation testing identifies 1,740 statistically significant SNVs, which are used in a second Random Forest model (NCRAD RF-RF), yielding a tuned model with *mtry*=83 and *ntree*=1,000, achieving 72.1% accuracy as shown in Figure 18 and identifying 581 significant SNVs through permutation testing.

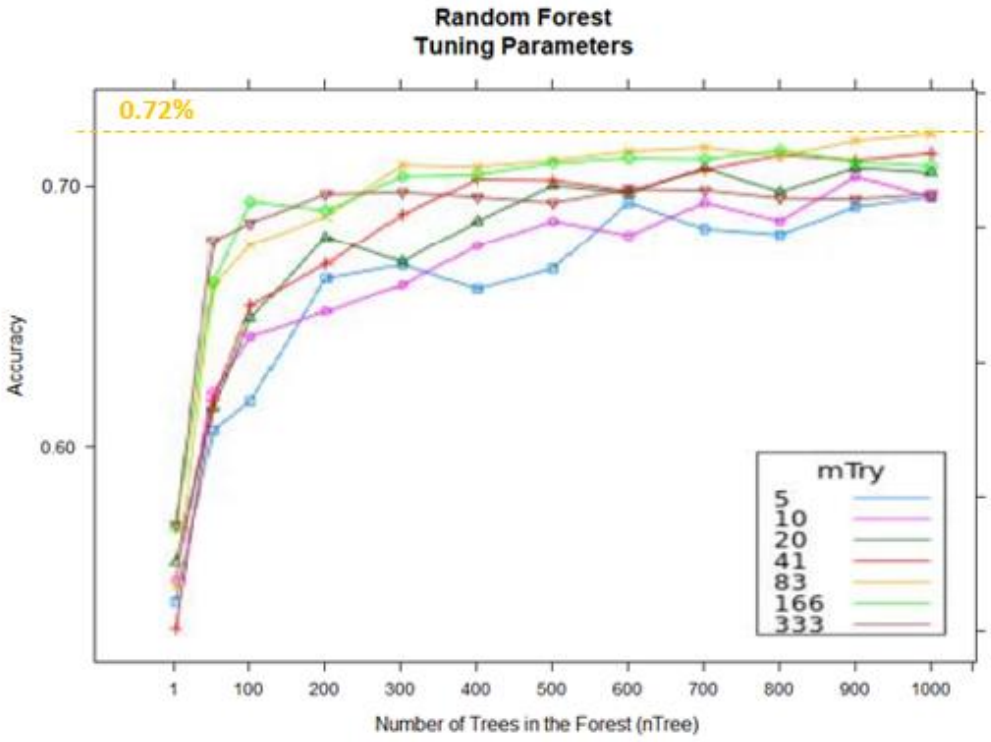


Figure 18: NCRAD RF-RF Tuning Parameters

For the ADNI RF-RF approach, the analysis of the ADNI dataset, which includes 7,639 significant SNVs, results in a final tuned model with *mtry*=43 and *ntree*=50, achieving an accuracy of 72.9% as shown in Figure 19.

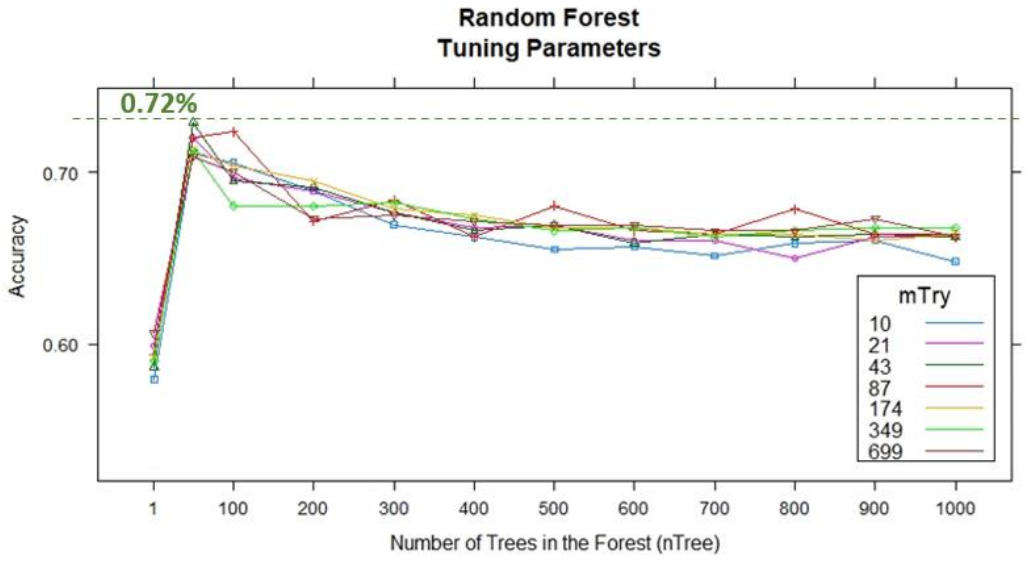


Figure 19: ADNI RF Tuning Parameters

Permuted significance testing was conducted with 100 permutations using the RANGER package. Permutation testing identifies 390 statistically significant SNVs, which are used in a second Random Forest model (ADNI RF-RF), yielding a tuned model with *mtry*=39 and *ntree*=50, achieving 74.0% accuracy as shown in Figure 20 and identifying 32 significant SNVs through permutation testing.

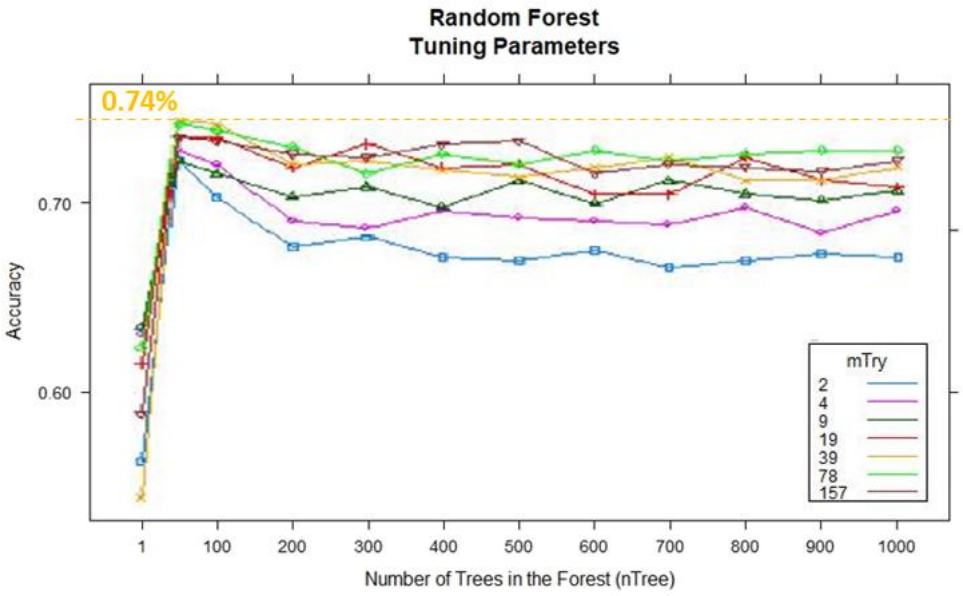


Figure 20: ADNI RF-RF Tuning Parameters

The resulting final RF parameters are given in Table 2.

Table 2: Tuning Parameters for Multi-Platform Genome Datasets

Dataset	Approach	mtry	ntree	Accuracy (%)
GenADA	LOAD RF	122	1000	92.44%
	LOAD RF-RF	2	900	85.08%
NCRAD	LOAD RF	1024	1200	68.80%
	LOAD RF-RF	83	1000	72.10%
ADNI	LOAD RF	43	50	72.90%
	LOAD RF-RF	39	50	74.00%

Figure 21 summarizes the number of significant single nucleotide variants obtained after applying multimodel Random Forest algorithms to three different random datasets by using the tuning parameters in Table 2.

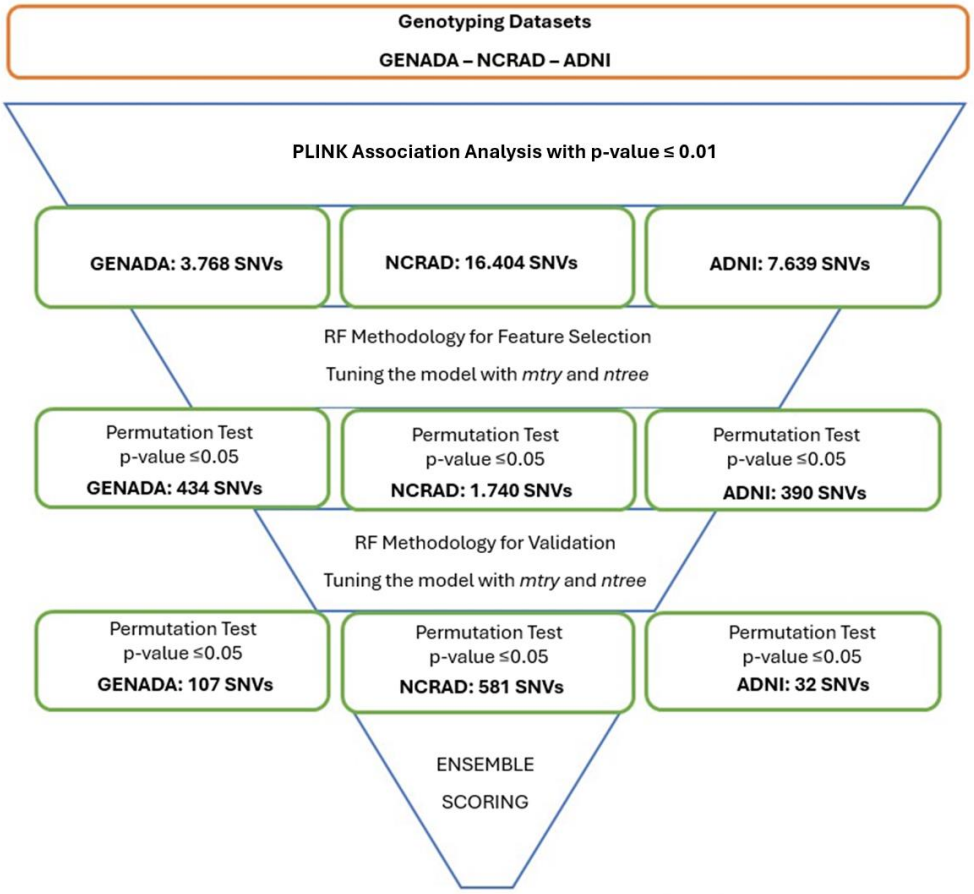


Figure 21: Number of Significant SNVs Selected with the Multimodel Approach (LOAD RF-RF)

Ensemble scoring will be performed for these variants obtained from individually validated models specific to the individual datasets at the cytogenetic location if there are no SNVs in common.

5.4. Intersection Evaluation of Selected SNVs

The individual LOAD RF models select 390 SNVs from the ADNI dataset, 1,740 from NCRAD, and 434 from GenADA, considering the permutation importance of variants at a 95% confidence level. No consensus variants were identified from three different datasets [9], as seen in Figure 7.

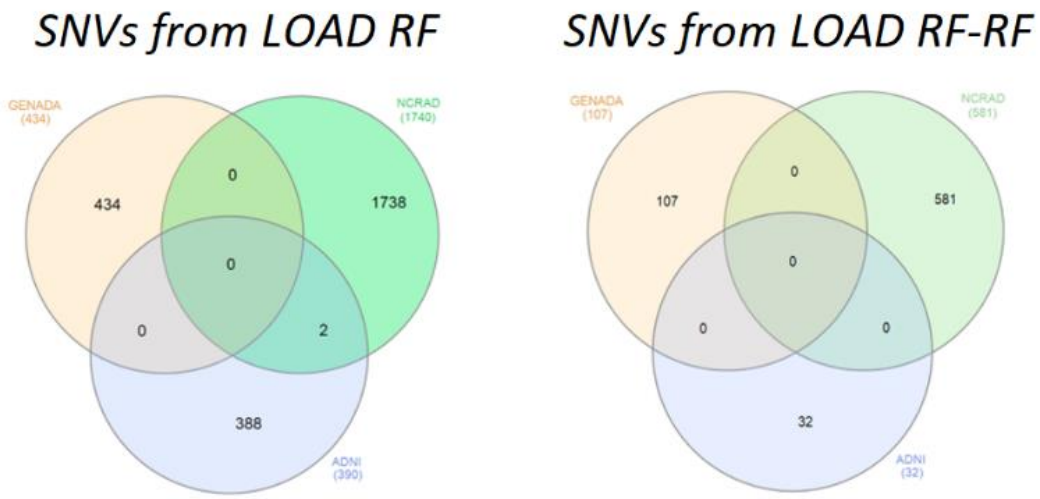


Figure 22: Number of Common SNVs from Three Individual Models

The RF-RF analysis identified the most significant SNVs. These RF-RF-identified SNVs are utilized during ensemble scoring to enrich the prioritized SNV list for causative variants since there is no in common.

5.5. Multi-Platform Variant Prioritization by EnSCAN Algorithm

We developed a scoring framework that maps individual SNVs to chromosomal position on genes, to observe nearby variations from various datasets because tagged SNVs do not overlap across platforms.

SNV annotation is determined using SNPnexus [54]. The pairwise Linkage Disequilibrium (LD) is determined by SNiPA [56]. The presence of the causal SNV in LD should be indicated by the correlation between the marker allele and the disease. Finding all of the LD blocks inside the genomic region adds to the knowledge of potential disease-causing variations [57]. The LD correlation coefficient is determined during the scoring process. The default value is entered as 0.7. LDs of each SNV are defined and labeled. Finally, the Sliding Window Algorithm (SWA) method of scoring

is applied to all SNVs found in the LOAD RF-RF model across three platforms that have multi-step models to prioritize them. The ensemble algorithm will evaluate each SNV within an iterated scoring window, resulting in three distinct scoring scenarios for each variant. Specifically, for each SNV, scores will be computed when the variant is positioned at the center, at the top, and the bottom of the window.

Each variant is calculated using the Ensemble Scoring Algorithm (EnSCAN) using the algorithm given in APPENDIX A, and those with the highest scores are selected. Figure 8 illustrates the distribution of these calculated scores.

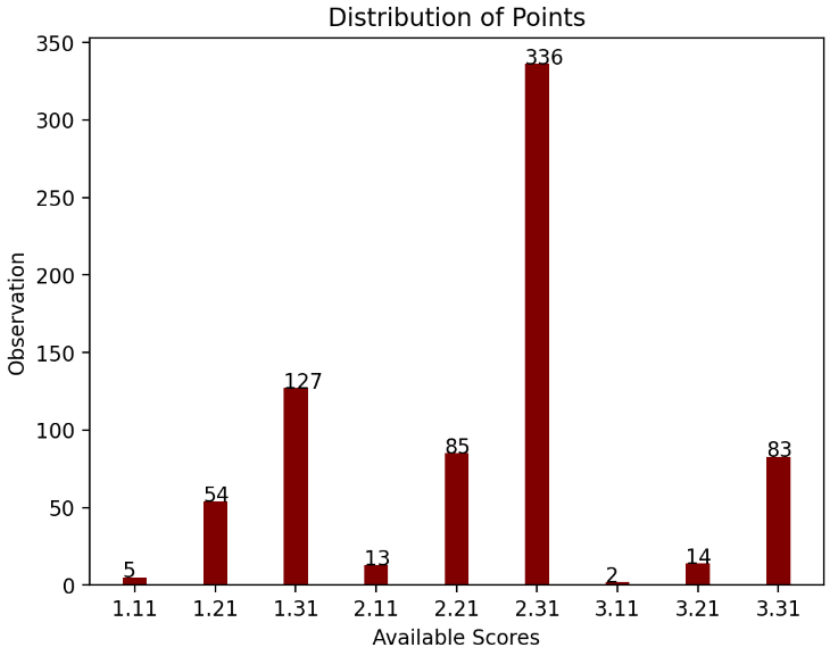


Figure 23: Ensemble Scoring Results

83 SNVs are scored with the highest EnSCAN score of "3.31". 43 out of 83 SNVs are identified as protein-coding as seen in APPENDIX B with details. Genes carrying these LOAD associated variants with highest EnSCAN score are CSMD2, NR5A2, KIF26B, CCDC3, CSGALNACT2, SLC18A2, PPFIBP2, TSPAN18, ALDH3B1, CNTN5, RNFT2, FBXW8, RBFOX1, NDRG4, PITPNC1, NPLOC4, PHLPP1, RUNX1, OSBP2, CYB5R3, LSAMP, TPRG1, FGF12, TENM3, ENPP6, ADCY2, ANKRD33B, IQGAP2, SPOCK1, ARHGAP26, CYFIP2, SYCP2L, SLC17A2, MDFI, RUNX2, CRISP1, PLAGL1, GNGT1, DLC1, C9orf3, SNX30, HMCN2 using GAD database in SNPnexus. Two different SNVs (rs2273570, rs9425897) are mapped to CSMD2. Among the 43 protein-coding variants with a score of 3.31, twelve mapped on chromosomes 5 and 6, located on ADCY2, ANKRD33B, IQGAP2, SPOCK1, ARHGAP26, CYFIP2, SYCP2L, SLC17A2, MDFI, RUNX2, CRISP1 and PLAGL1 genes.

Also, 29 out of 43 variants on protein-coding genes were associated with known disease phenotypes. Among these, 6.5% were directly linked to Alzheimer's Disease (AD),

while 35.5% were associated with other AD-related conditions, including cholesterol metabolism, Type 2 Diabetes, cardiovascular disorders, and immunological issues. Furthermore, 30.5% were tied to addiction-related phenotypes such as Tobacco Use Disorder. Additionally, 13% were associated with various neuropsychiatric disorders, and the remaining 14.5% were linked to non-neurological or cancer-related phenotypes.

5.6. Evaluation of Ensemble Scores

We conduct functional enrichment analysis, including over-representation analysis (ORA) and gene set enrichment analysis, to validate the EnSCAN variants across different scoring categories. This involves mapping the single nucleotide variants (SNVs) to genes to access established functional information sources and identify statistically significant enriched terms. This approach allows us to pinpoint effective enrichments that are related to Alzheimer's disease.

Given that EnSCAN class scores are discrete, representing distinct SNV categories rather than continuous values, it is not advisable to set a fixed threshold. Each class reflects different characteristics of the selected SNVs, and applying a rigid threshold could obscure valuable disease-related information. Thus, it is crucial to consider the unique properties of each class when interpreting results. For example, a score of 3.31 represents the highest level of significance, indicating strong support across all datasets and robust associations with the disease.

This score range reflects the degree of association between each SNV and the disease, incorporating information from multiple datasets and analytical models. The 3.31 score class includes SNVs from all three datasets, selected from independent linkage disequilibrium (LD) regions, making them the most prioritized variants. Lower score classes, on the other hand, suggest weaker disease associations. Enrichment analyses have been performed for each score to identify molecular pathways linked to Alzheimer's disease (AD). Depending on the purpose of the downstream analysis, different score classes can be employed. High-scoring variants, such as those with a score of 3.31, maybe prioritized for experimental validation with independent case-control studies, while lower-scoring variants, like those with a score of 2.31, are more suitable for the enrichment and predictive analyses, offering a broader view of the studied conditions.

In our study, we focus on identifying pathways potentially involved in the development of late-onset Alzheimer's disease (LOAD) and examine how specific scores align with biological mechanisms. This analysis enhances our understanding of the biological context of the findings and helps pinpoint potential targets for therapeutic intervention.

To manage a large number of variants in certain EnSCAN score classes and present the results effectively, we used an evaluation method that supports visualization and simplifies interpretation. Enrichment analysis based on Gene Ontologies and Reactome Pathways was chosen to reveal the molecular and disease-level connections between the selected variants.

The pathways from the following EnSCAN score categories are benchmarked against all calculated scores:

1) All variants with EnSCAN class scores as a benchmark
2) Variants with an EnSCAN class score equal to 2.31
3) Variants with EnSCAN class scores of 2.31 or 3.31
4) Variants with EnSCAN class scores of 3.11, 3.21, or 3.31 (3 s)
5) Variants with EnSCAN class scores equal to 3.31

5.6.1. Functional Evaluation of All LOAD RF-RF Variants

When SNVs with EnSCAN class scores were evaluated, enrichment analysis revealed a general view of the biological process possibly underlying the molecular etiology of AD as illustrated in Figure 24.

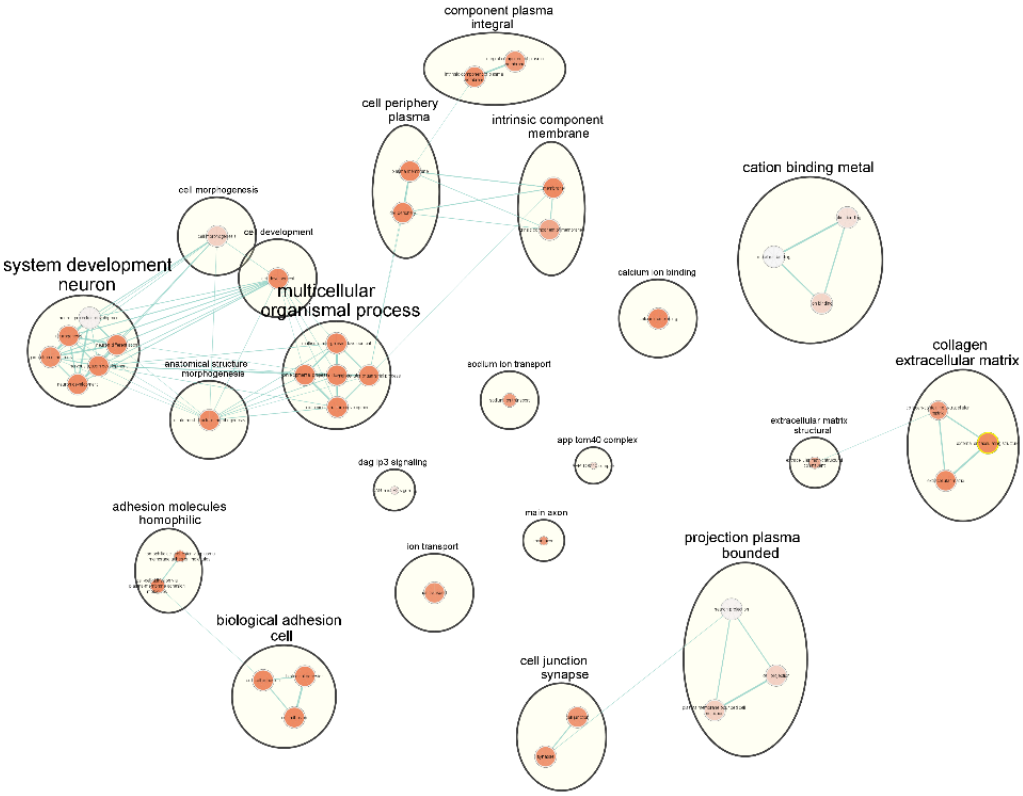


Figure 24: Functional Enrichment Analysis of All Variants with EnSCAN Class Scores

Tissue and cell-type specificity are fundamental to understanding human physiology and disease, particularly within the context of the "multicellular organismal process" [58]. Investigating the genetic underpinnings of complex tissues and distinct cell lineages is crucial for advancing both diagnostic and therapeutic approaches.

Exploring tissue-specific networks offers a novel framework for generating hypotheses regarding the molecular mechanisms underlying human diseases.

The development of the neuronal system is also integral to the study of Alzheimer's Disease (AD). Efforts to mitigate symptoms associated with aging and age-related conditions, such as AD, have included various strategies aimed at stimulating neurogenesis, through both pharmacological and natural methods. Research into the regulatory mechanisms of stem cell neurogenesis and the functional integration of newborn neurons has laid the groundwork for stem cell therapy. A review [59] provides a comprehensive overview of AD pathology across different neural and glial cell types, and it summarizes current experimental strategies involving stem cell treatments and their prospective applications in clinical practice.

A study [60] shows that abnormal "ion channels" (especially those showing selectivity for cations like Ca2+) might be a reason for the toxic mechanism behind AD pathology. "Peptides or phosphopeptides" of common plasma proteins show increased observation frequency by Chi-Square and/or precursor intensity in AD, according to a recent study [61]. Increases in mean precursor intensity of peptides from common plasma proteins such as DISC1, EXOSC5, UBE2G1, SMIM19, NXNL1, PANO, EIF4G1, KIR3DP1, MED25, MGRN1, OR8B3, MGC24039, POLR1A, SYTL4, RNF111, IREB2, ANKMY2, SGKL, SLC25A5, CHMP3 among others were observed in association with AD.

Disruptions in the "extracellular matrix" [62] have been implicated in the pathology of Alzheimer's Disease (AD), with effects on various processes including synaptic transmission, amyloid-β plaque formation and clearance, Tau protein production, oxidative stress response, and inflammatory reactions. The extracellular matrix is composed of numerous macromolecules secreted by cells, including collagen, elastin, fibronectin, laminin, and glycoproteins. Modulating the expression of these matrix components is crucial for stabilizing or potentially ameliorating the progression of the disease. For instance, collagen fibrils can form around susceptible neurons, obstructing the binding of amyloid-β (Aβ) proteins to the cells. Consequently, components of the extracellular matrix represent potential targets and biomarkers for the development and treatment of AD.

5.6.2. Functional Evaluation of EnSCAN Class Score 2.31

When SNVs whose scores are 2.31 with an EnSCAN class score were evaluated, the enrichment analysis revealed a general view of the biological process possibly underlying the molecular etiology of AD as illustrated in Figure 25.

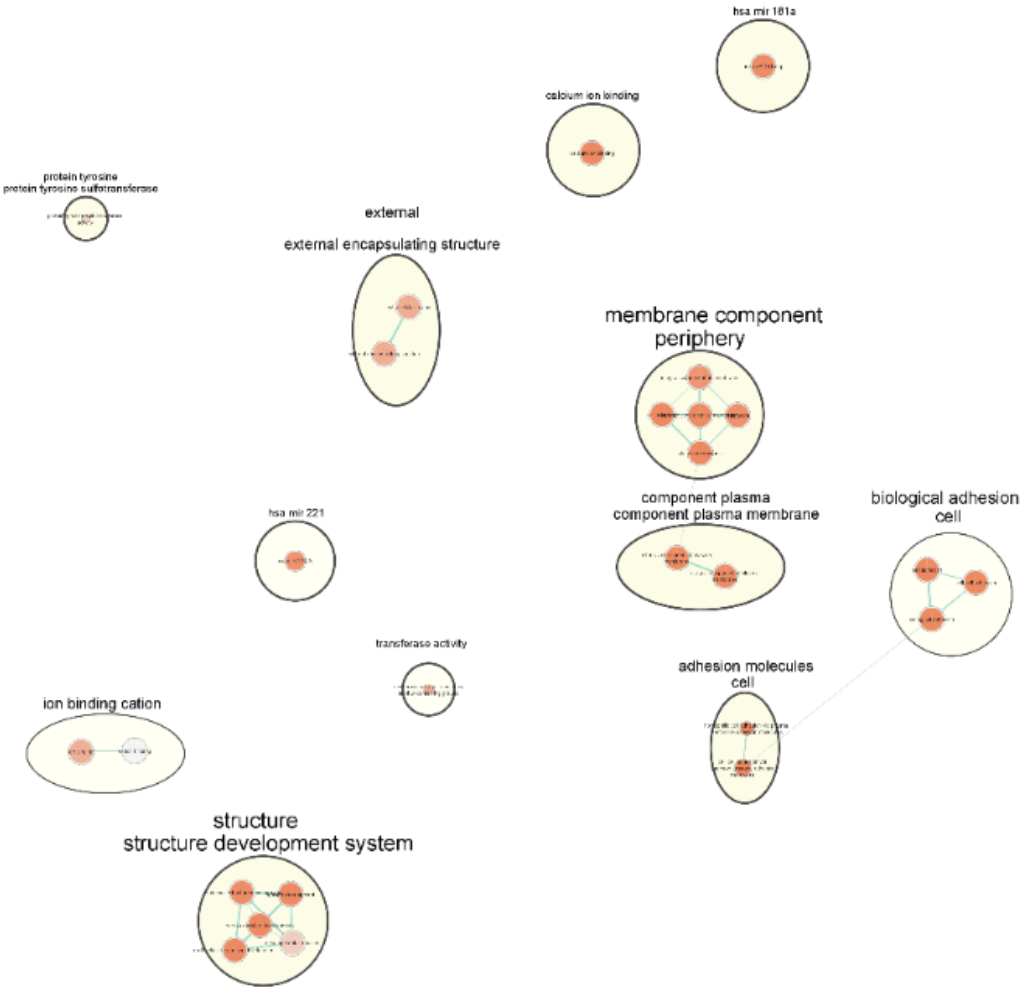


Figure 25: Functional Enrichment Analysis of Variants with EnSCAN Class Score Equal to 2.31

Enrichment analysis of only SNVs with an EnSCAN score of 2.31 showed that AD-related processes of "adhesion molecules/biological cell adhesion", "cation ion binding/calcium ion binding," and "external encapsulating structure/plasma membrane/periphery membrane component" are enriched within this category.

5.6.3. Functional Evaluation of Union of Class Score 2.31 and 3.31

When SNVs whose scores are 2.31 and 3.31 with EnSCAN were evaluated, the enrichment analysis revealed a general view of the biological process possibly underlying the molecular etiology of AD as illustrated in Figure 26.

SNVs with an EnSCAN score of 2.31 or 3.31 were evaluated together under one category; the top enrichment classes were not changed compared to only 2.31 and the benchmark.

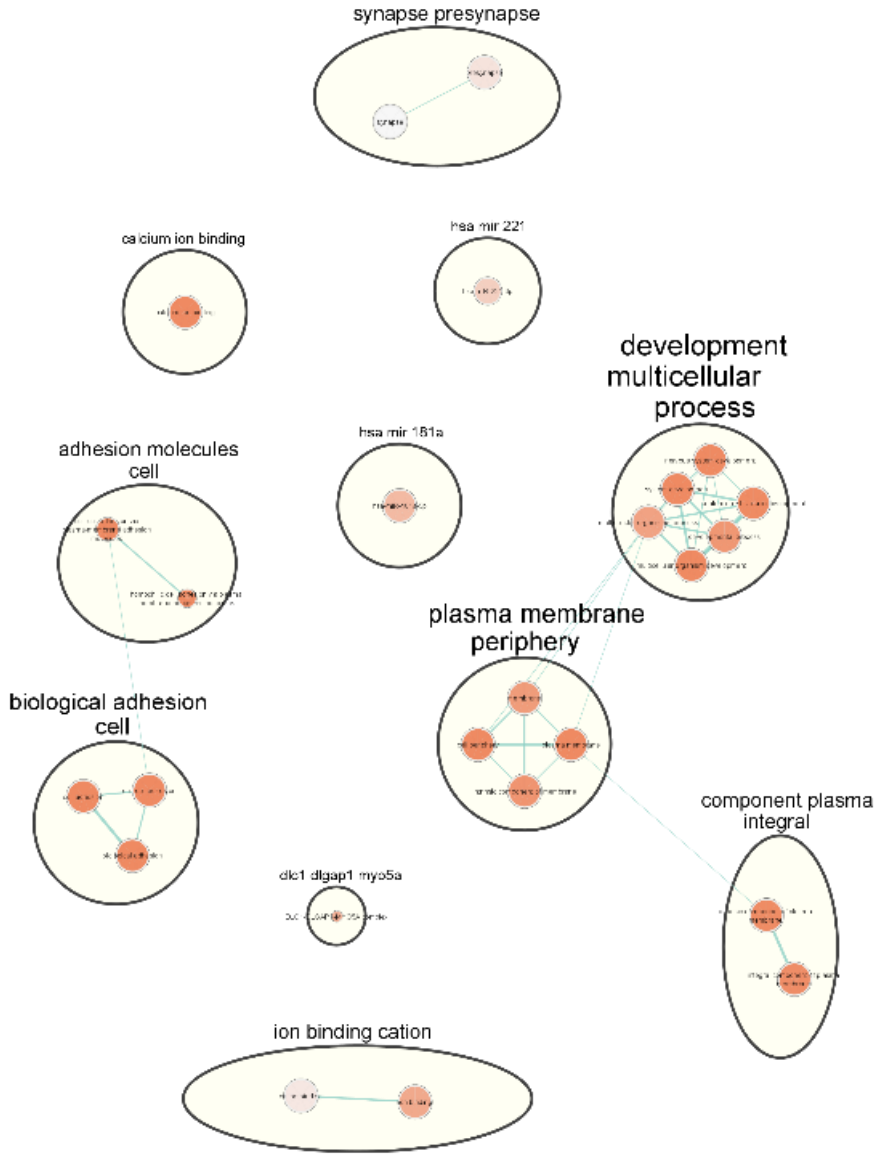


Figure 26: Functional Enrichment Analysis of Union of Variants with EnSCAN Class Score 2.31 or 3.31

From an ontological perspective, several significant terms influencing susceptibility to Alzheimer's Disease in humans have been identified. Considering ontology, development multicellular process, external encapsulating structure, has mir 221, dlc1 dlgap1 myo5a, metal ion transmembrane, plasma membrane periphery, calcium regulation cardiac, component plasma integral, biological adhesion cell, synapse presynapse, adhesion molecules cell, has mir181a, calcium ion binding and cation binding ion are significant terms that affect the susceptibility to Alzheimer's Disease in humans.

All three categories analyzed in Figure 26 for the enriched biological process shared common keywords of "adhesion molecules/biological cell adhesion", "ion cation

binding/calcium ion binding", and "plasma membrane/ peripheral component/ plasma integral". One biological annotation category that emerges in this set different from 2.31's enrichment analysis that is noteworthy is "synapses and presynapses". The enrichment analysis allowed us to observe biological processes involving LOAD-associated variants. Comparing EnSCAN scoring categories also revealed the consistent enrichment of ontological terms between categories that are known to have roles in the core of AD pathophysiology.

5.6.4. Functional Evaluation of the Union of EnSCAN Class Score 3.11, 3.21, 3.31 (all 3s):

When SNVs whose scores are 3s (3.11, 3.21 and 3.31) with EnSCAN were evaluated, the enrichment analysis revealed a general view of the biological process possibly underlying the molecular etiology of AD as illustrated in Figure 27.

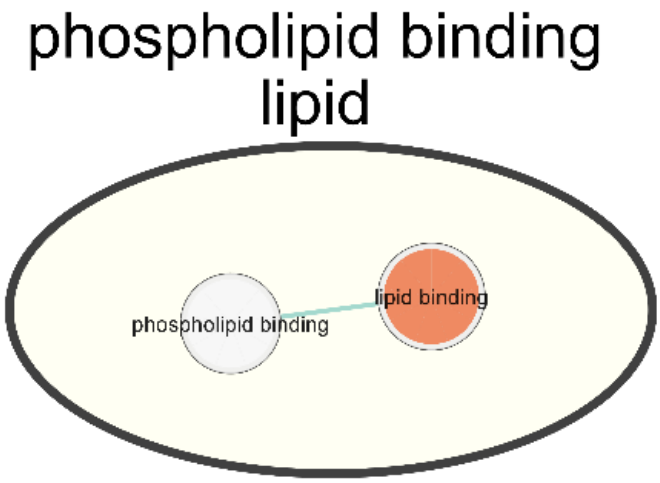


Figure 27: Functional Enrichment Analysis of Variants with EnSCAN Class Score 3.11 or 3.21 or 3.31 (3s)

Focusing on neighboring LOAD RF-RF variants identified in only three datasets with EnSCAN class scores of 3.11, 3.21, and 3.31, we found that the enrichment analysis was primarily limited to "phospholipid binding" as seen in Figure 27, a topic that has gained attention in AD pathophysiology. Abnormal phospholipid binding is thought to contribute to the development of Alzheimer's disease (AD), with mutations in genes involved in phospholipid metabolism being associated with an increased AD risk [63]. Although further research is required to clarify the role of phospholipid interactions in AD, recent studies suggest that certain phospholipids may encourage the formation of alpha-synuclein aggregates [64], which could disrupt normal phospholipid interactions within cell membranes.

5.6.5. Functional Evaluation of Class Score 3.31

When SNVs whose scores are 3.31 with EnSCAN were evaluated, the enrichment analysis revealed a general view of the biological process possibly underlying the molecular etiology of AD as illustrated in Figure 28.

Figure 28: Enrichment Analysis of Variants with EnSCAN Class Score Equal to 3.31

The enrichment analysis of variants with top EnSCAN scores reveals the "nicotinic activity on dopaminergic neurons" as the only ontology common between these variants as seen in Figure 28.

The systematic review [65] and meta-analysis are the first to investigate the relationship between Alzheimer's disease (AD) and both dopamine levels and dopamine receptors. It highlights that AD is linked to brain damage from amyloid plaques and neurofibrillary tangles, which disrupt synaptic function and neurotransmission. The review consolidates evidence from genetic, experimental, and clinical studies suggesting that dopaminergic dysfunction, including reduced dopamine levels and decreased D2-like receptors in the hippocampus and frontal cortex, is associated with AD. Additionally, dopamine's role in modulating cortical excitability and cholinergic transmission points to potential therapeutic strategies, such as dopamine D2 agonists like rotigotine, which may improve cognitive function in AD. Generally, the activation of Nicotinic Acetylcholine Receptors (nAChRs) on dopaminergic neurons enhances dopamine release, which in turn influences various cognitive functions, such as memory, attention, and reward processing [65]. Dopamine is known to have a role in cognitive functions, and decreased dopamine levels are reported [65] as a feature of AD. However, the interplay between nicotinic activity on dopaminergic neurons and AD is complex, and the specific impacts of nicotine on dopaminergic neurons, particularly regarding cognitive decline, memory impairment, and AD progression, remain under active investigation [66].

In a study [67], nicotine has been shown to offer cognitive benefits across various contexts, including in individuals with cognitive impairments. Research indicates that nicotine can enhance cognitive function in Alzheimer's disease (AD) by targeting specific molecular pathways. In AD, nicotine improves cognitive deficits by boosting Akt activity and stimulating the phosphoinositide 3-kinase/Akt signaling pathway [67], which supports learning and memory.

Nicotine administration has shown the potential to enhance memory in healthy individuals, particularly when memory impairments are caused by sleep deprivation.

Sleep deprivation negatively affects memory by reducing the phosphorylation of CaMKII, a key regulator of cell proliferation and synaptic plasticity. CaMKII is crucial for the expression and trafficking of glutamate receptors, which are essential for memory and brain function [67]. Research indicates that acute nicotine use can counteract sleep deprivation-induced memory deficits by increasing CaMKII phosphorylation, thereby improving memory performance in these scenarios. However, while these findings highlight nicotine's potential cognitive benefits, more research is needed to fully understand the impact of both acute and chronic nicotine use on memory function.

CHAPTER 6

DISCUSSION

The progression of high-throughput genotyping and next-generation sequencing technologies has significantly enhanced the ability to conduct genetic epidemiological studies on large-scale datasets. These advancements have facilitated the identification of single nucleotide variations (SNVs) associated with a range of complex diseases. Numerous research institutions gather and share Genome-Wide Association Study (GWAS) data to enable more in-depth analyses, aiming to shed light on the genetic underpinnings of various diseases. While the integration or meta-analysis of GWAS data from different centers has led to the discovery of variants strongly linked to these conditions, pinpointing causative variants remains a challenging task.

This thesis focuses on the development of an "ensemble" model that has not been previously utilized in literature. The model has been designed to provide an innovative contribution to existing techniques in machine learning. Ensemble methods aim to enhance overall performance by combining predictions from multiple models, often resulting in improved accuracy and generalization capabilities. In this context, the developed model has demonstrated its effectiveness in high-dimensional genetic datasets and its ability to capture complex relationships more effectively. The significance of this model in the field of machine learning lies in its capacity to enable a more thorough analysis of data and a more comprehensive examination of genetic variations. This novel approach transcends current methods, offering the potential for more precise and reliable outcomes.

Our main goal was to develop a methodology for ensemble analysis of GWAS data from different centers and genotyping platforms to reveal associated cytogenetic loci and focus on variants with a higher probability of being causative. To integrate multi-platform data, we have investigated machine learning methods for knowledge extraction. RF-RF models are trained for each dataset of Alzheimer's disease. An assembly of prioritized variants selected through RF-RF models is formed by mapping variants selected in each RF-RF model to their cytogenetic loci. The EnSCAN scoring system is proposed to survey all selected variants and categorize variants depending on whether they are on a locus independently selected in all GWAS datasets analyzed. The core of the presented study was the ensemble of priori information derived from

three different models containing different representative variants, along with the newly proposed EnSCAN algorithm. The enrichment analysis for variants based on their EnSCAN score showed how biological processes and annotations are narrowed down as the EnSCAN score increases, allowing researchers to focus their studies on enriched biological processes through a smaller associated variants list.

In this study, we introduce a novel approach to uncovering hidden patterns by integrating results from various multi-platform genotyping studies. We utilized the Ensemble Scoring algorithm to identify single nucleotide variants (SNVs) that serve as biomarkers for the early or differential diagnosis of Alzheimer's disease. This was achieved by developing machine learning models that highlight significant patterns, accounting for the complex interactions between variants. The application of these models revealed critical biomarkers associated with Alzheimer's disease, as determined through the Ensemble Scoring algorithm.

EnSCAN scoring is proposed to prioritize LOAD RF-RF variants neighboring with LD independent variants in its cytogenetic loci from two or three datasets (EnSCAN score of 2.31 or 3.31). This approach effectively narrows down the list of LOAD-associated variants while preserving the representation of relevant biological processes. Therefore, for future ensemble analyses of multi-platform GWAS data, variants with EnSCAN scores of 2.31 and 3.31 should be considered for in-depth investigation to identify potential causative variants among the associated single nucleotide variants (SNVs).

A total of 85 variants are identified with the LOAD RF-RF model, with the top score of 3.31, and 29 of the variants on protein-coding genes were shown to be associated with known disease phenotypes, including Alzheimer's Disease (AD) and other AD-related conditions such as cholesterol metabolism, Type 2 Diabetes, cardiovascular disorders, and immunological issues. Besides various neuropsychiatric disorders and other disease phenotypes, interestingly, there was a strong representation of Tobacco Use Disorder.

APOE gene variants are known to increase the predictive power. Research [12] indicates that the presence of APOE ε4 or ε2 alleles independently influences the differential diagnosis of AD, particularly in early-onset and familial cases. However, a limitation of our study is that genotyping data for the APOE gene was not available across all three datasets (NCRAD, ADNI, GenADA). As a result, our analysis did not include APOE allele information, preventing us from assessing their potential impact on late-onset Alzheimer's disease (LOAD) or their interactions with the variants identified in this study.

While we have demonstrated associations with Alzheimer's Disease through enrichment analysis and disease phenotypes, it is essential to investigate the SNVs identified by the LOAD RF-RF model within a clinical framework. This examination aims to assess the predictive efficacy of EnSCAN scores for early or differential

diagnosis of LOAD. The discovery of new associated and causative variants for AD through LOAD RF-RF modeling and EnSCAN scoring can lead to the development of cost- and time-effective diagnoses and an understanding of pathophysiology leading to neurodegeneration. Such studies will accelerate the new studies focusing on early and differential diagnosis of AD and developing new therapeutics. Diagnosing and treating such a high-burden disease in the early phases will eventually bring significant social and economic benefits.

Overall, the novelty of our approach is based on the assembly of multi-platform GWAS data along with the proposed EnSCAN scoring of variants. Proposed EnSCAN scoring, applied to variants identified through machine learning modeling, can unveil concealed, novel, and meaningful patterns by considering non-linear interactions between variants. This approach enhances our comprehension of the genetic predisposition in complex genetic disorders, where the cumulative impact of multiple variants determines the risk.

During this study, several risks and limitations were identified, which could have influenced the results. One major risk was that the proposed multimodal machine learning approach might not effectively contribute to clinical decision support, which was mitigated by continuously monitoring the literature and conducting pathway analysis. Another challenge was the potential incompatibility of the three models, addressed by mapping SNVs to cytogenetic locations to facilitate model integration. Additionally, the application of graph theory for model minimization was initially inefficient, but was improved through evidence-based minimization, removing less significant nodes. The study also faced computational challenges due to the high throughput of genomic data, which was overcome by utilizing high-performance computing infrastructure. Lastly, the variability in sequencing platforms was anticipated and managed through the development of an innovative ensemble scoring approach.

Despite these challenges, the mitigation strategies ensured the study's success and provided valuable insights for future research in genomic data analysis. This framework demonstrates considerable potential as a robust method for post-GWAS analysis, particularly for exploring complex genetic disorders.

REFERENCES

[1] X. Mao, B. D. Young, and Y.-J. Lu, "The application of single nucleotide polymorphism microarrays in cancer research.," *Curr. Genomics*, vol. 8, no. 4, pp. 219–28, 2007, doi: 10.2174/138920207781386924.

[2] C. Reitz, E. Rogaeva, and G. W. Beecham, "Late-onset vs nonmendelian early-onset Alzheimer disease," *Neurol. Genet.*, vol. 6, no. 5, p. e512, 2020, doi: 10.1212/nxg.0000000000000512.

[3] E. Bagyinszky, Y. C. Youn, S. S. A. An, and S. Kim, "The genetics of Alzheimer's disease," *Clinical Interventions in Aging*, vol. 9, no. July. pp. 535–551, 2014, doi: 10.2147/CIA.S51571.

[4] T. P. V. Huynh, A. A. Davis, J. D. Ulrich, and D. M. Holtzman, "Apolipoprotein E and Alzheimer's disease: The influence of apolipoprotein E on amyloid-β and other amyloidogenic proteins," *J. Lipid Res.*, vol. 58, no. 5, pp. 824–836, 2017, doi: 10.1194/jlr.R075481.

[5] S. S. Muñoz, B. Garner, and L. Ooi, "Understanding the Role of ApoE Fragments in Alzheimer's Disease," *Neurochem. Res.*, vol. 44, no. 6, pp. 1297–1305, 2019, doi: 10.1007/s11064-018-2629-1.

[6] N. Ertekin-Taner, "Genetics of Alzheimer disease in the pre- and post-GWAS era.," *Alzheimers. Res. Ther.*, vol. 2, no. 1, p. 12, 2010, doi: 10.1186/alzrt26.

[7] K. Bettens, K. Sleegers, and C. Van Broeckhoven, "Current status on alzheimer disease molecular genetics: From past, to present, to future," *Hum. Mol. Genet.*, vol. 19, no. R1, pp. 4–11, 2010, doi: 10.1093/hmg/ddq142.

[8] B. Duthey, "Background Paper 6.11 Alzheimer Disease and other Dementias, Update on 2004," *World Heal. Organ.*, no. February, pp. 1–77, 2013.

[9] S. Jayadev *et al.*, "Alzheimer's disease phenotypes and genotypes associated with mutations in presenilin 2," *Brain*, vol. 133, no. 4, pp. 1143–1154, 2010, doi: 10.1093/brain/awq033.

[10] K. Sleegers *et al.*, "A 22-single nucleotide polymorphism Alzheimer's disease risk score correlates with family history, onset age, and cerebrospinal fluid A??42," *Alzheimer's Dement.*, vol. 11, no. 12, pp. 1452–1460, 2015, doi: 10.1016/j.jalz.2015.02.013.

[11] C. R. Jack and D. M. Holtzman, "Biomarker modeling of alzheimer's disease," *Neuron*, vol. 80, no. 6, pp. 1347–1358, 2013, doi: 10.1016/j.neuron.2013.12.003.

[12] M. Kikuchi *et al.*, “Polygenic effects on the risk of Alzheimer’s disease in the Japanese population,” *medRxiv*, p. 2023.10.06.23296656, 2023, [Online]. Available: https://www.medrxiv.org/content/10.1101/2023.10.06.23296656v1.

[13] Y. Li and L. Chen, “Big Biological Data: Challenges and Opportunities Expanding volume of the big biological data and its bonanza,” *Genomics. Proteomics Bioinformatics*, vol. 12, no. 5, pp. 187–189, 2014, doi: 10.1016/j.gpb.2014.10.001.

[14] V. Botta, G. Louppe, P. Geurts, and L. Wehenkel, “Exploiting SNP correlations within random forest for genome-wide association studies,” *PLoS One*, vol. 9, no. 4, 2014, doi: 10.1371/journal.pone.0093379.

[15] S. J. Winham *et al.*, “SNP interaction detection with Random Forests in high-dimensional genetic data.,” *BMC Bioinformatics*, vol. 13, no. 1, p. 164, 2012, doi: 10.1186/1471-2105-13-164.

[16] M. ÇOLAK, T. TÜMER SİVRİ, N. PERVAN AKMAN, A. BERKOL, and Y. EKİCİ, “Disease prognosis using machine learning algorithms based on new clinical dataset,” *Commun. Fac. Sci. Univ. Ankara Ser. A2-A3 Phys. Sci. Eng.*, vol. 65, no. 1, pp. 52–68, 2023, doi: 10.33769/aupse.1215962.

[17] S. C. Yücebaş, “A Hybrid Feature Selection Model For Genome Wide Association Studies,” Middle East Technical University, 2013.

[18] A. L. Tarca, V. J. Carey, X. Chen, R. Romero, and S. Drăghici, “Machine learning and its applications to biology.,” *PLoS Comput. Biol.*, vol. 3, no. 6, p. e116, 2007, doi: 10.1371/journal.pcbi.0030116.

[19] R. Jiang, W. Tang, X. Wu, and W. Fu, “A random forest approach to the detection of epistatic interactions in case-control studies.,” *BMC Bioinformatics*, vol. 10 Suppl 1, p. S65, 2009, doi: 10.1186/1471-2105-10-S1-S65.

[20] H. Kashyap, H. A. Ahmed, N. Hoque, S. Roy, and D. K. Bhattacharyya, “Big Data Analytics in Bioinformatics: A Machine Learning Perspective,” *J. Latex Cl. Files*, vol. 13, no. 9, pp. 1–20, 2014, [Online]. Available: http://arxiv.org/abs/1506.05101.

[21] M. N. Wright and A. Ziegler, “Ranger: A fast implementation of random forests for high dimensional data in C++ and R,” *J. Stat. Softw.*, vol. 77, no. 1, pp. 1–17, 2017, doi: 10.18637/jss.v077.i01.

[22] O. Erdoğan, “Predicting the Disease Alzheimer (AD) With SNP Biomarkers and Clinical Data Based Decision Support System Using Data Mining Classification Approaches,” METU, 2012.

[23] G. Tosto and C. Reitz, “Genome-wide association studies in Alzheimer’s disease: a review.,” *Curr. Neurol. Neurosci. Rep.*, vol. 13, no. 10, p. 381, 2013, doi: 10.1007/s11910-013-0381-0.

[24] B. S. Shastry, “SNP alleles in human disease and evolution.,” *J. Hum. Genet.*, vol. 47, no. 11, pp. 561–566, 2002, doi: 10.1007/s100380200086.

[25] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[26] B. Leo, F. Jerome, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. 1984.

[27] L. Breiman, “Out-of-bag estimation,” pp. 1–13, 1996, doi: 10.1016/j.patcog.2009.05.010.

[28] B. A. Goldstein, E. C. Polley, and F. B. S. Briggs, “Random forests for genetic association studies.,” *Stat. Appl. Genet. Mol. Biol.*, vol. 10, no. 1, p. 32, 2011, doi: 10.2202/1544-6115.1691.

[29] K. L. Lunetta, L. B. Hayward, J. Segal, and P. Van Eerdewegh, “Screening large-scale association study data: exploiting interactions using random forests.,” *BMC Genet.*, vol. 5, no. 1, p. 32, 2004, doi: 10.1186/1471-2156-5-32.

[30] S. J. Winham *et al.*, “SNP interaction detection with Random Forests in high-dimensional genetic data,” *BMC Bioinformatics*, vol. 13, no. 1, p. 164, 2012, doi: 10.1186/1471-2105-13-164.

[31] N. Briones and V. Dinu, “Data mining of high density genomic variant data for prediction of Alzheimer’s disease risk,” *BMC Med. Genet.*, vol. 13, no. 1, p. 7, 2012, doi: 10.1186/1471-2350-13-7.

[32] E. E. Tripoliti, D. I. Fotiadis, M. Argyropoulou, and G. Manis, “A Six Stage Approach for the Diagnosis of the Alzheimer’s Disease Based on fMRI Data.,” *J. Biomed. Inform.*, vol. 43, no. 2, pp. 307–20, Apr. 2010, doi: 10.1016/j.jbi.2009.10.004.

[33] J. N. Hirschhorn and M. J. Daly, “Genome-wide association studies for common diseases and complex traits,” *Nat. Rev. Genet.*, vol. 6, no. 2, pp. 95–108, 2005, doi: 10.1038/nrg1521.

[34] Y. Jung, J. Z. Huang, and J. Hu, “Biomarker Detection in Association Studies: Modeling SNPs Simultaneously via Logistic ANOVA,” *J. Am. Stat. Assoc.*, no. May 2015, pp. 00–00, 2014, doi: 10.1080/01621459.2014.928217.

[35] V. Botta and S. Hansoul, “Prediction of genetic risk of complex diseases by supervised learning,” in *The annual machine learning conference of Belgium*

and The Netherlands, 2008, pp. 83–84.

[36] G. W. Beecham *et al.*, "Genome-Wide Association Meta-analysis of Neuropathologic Features of Alzheimer's Disease and Related Dementias," *PLoS Genet.*, vol. 10, no. 9, 2014, doi: 10.1371/journal.pgen.1004606.

[37] L. Bertram and R. E. Tanzi, "Genome-wide association studies in Alzheimer's disease," *Hum. Mol. Genet.*, vol. 18, no. R2, pp. 137–145, 2009, doi: 10.1093/hmg/ddp406.

[38] G. Üskünkar, "An Integrative Approach To Structured SNP Prioritization and Representative SNP Selection For Genome-wide Association Studies:Algorithms and Systems," Middle East Technical University, 2008.

[39] B. A. Goldstein, A. E. Hubbard, A. Cutler, and L. F. Barcellos, "An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings.," *BMC Genet.*, vol. 11, p. 49, 2010, doi: 10.1186/1471-2156-11-49.

[40] X. Chen and H. Ishwaran, "Random forests for genomic data analysis," *Genomics*, vol. 99, no. 6, pp. 323–329, 2012, doi: 10.1016/j.ygeno.2012.04.003.

[41] F. F. Sherif, N. Zayed, and M. Fakhr, "Discovering Alzheimer Genetic Biomarkers Using Bayesian Networks," *Adv. Bioinformatics*, vol. 2015, pp. 1–9, 2015, doi: 10.1155/2015/639367.

[42] W. Wei, S. Visweswaran, and G. F. Cooper, "The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data.," *J. Am. Med. Inform. Assoc.*, vol. 18, no. 4, pp. 370–375, 2011, doi: 10.1136/amiajnl-2011-000101.

[43] M. Vannucci, *Bayesian Models for Integrative Genomics*, no. January 2016. 2013.

[44] P. E. Spies *et al.*, "A prediction model to calculate probability of Alzheimer's disease using cerebrospinal fluid biomarkers.," *Alzheimers Dement.*, vol. 9, no. 3, pp. 1–7, 2012, doi: 10.1016/j.jalz.2012.01.010.

[45] L. L. Gu, R. Q. Yang, Z. Y. Wang, D. Jiang, and M. Fang, "Ensemble learning for integrative prediction of genetic values with genomic variants," *BMC Bioinformatics*, vol. 25, no. 1, pp. 1–18, 2024, doi: 10.1186/s12859-024-05720-x.

[46] B. Yaldız, O. Erdoğan, S. Rafatov, C. Iyigün, and Y. Aydın Son, "Revealing third-order interactions through the integration of machine learning and entropy methods in genomic studies," *BioData Min.*, vol. 17, no. 1, pp. 1–17, 2024, doi: 10.1186/s13040-024-00355-3.

[47] E. Mountjoy *et al.*, "An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci," *Nat. Genet.*, vol. 53, no. 11, pp. 1527–1533, 2021, doi: 10.1038/s41588-021-00945-5.

[48] I. Barrio-Hernandez *et al.*, "Network expansion of genetic associations defines a pleiotropy map of human cell biology," *Nat. Genet.*, vol. 55, no. 3, pp. 389–398, 2023, doi: 10.1038/s41588-023-01327-9.

[49] O. Erdoğan and Y. Aydin Son, "Predicting the disease of Alzheimer with SNP biomarkers and clinical data using data mining classification approach: decision tree.," in *Studies in health technology and informatics*, 2014, vol. 205, pp. 511–5, doi: 10.3233/978-1-61499-432-9-511.

[50] Y. Aydin Son, "Modelling of the Alzheimer's Disease Based on Genomic And Clinical Findings and Investigation of Polymorphisms Proposed in the Model on a Turkish Case Group," Ankara, 2020. [Online]. Available: https://avesis.metu.edu.tr/proje/ba9e7ea5-8ed2-4015-8932-fcce72a551ba/alzheimer-hastaliginin-erken-ve-ayirici-tanisi-icin-genomik-ve-klinik-bilgiye-dayali-in-silico-model-ve-genotiplendirme-kiti-gelistirilmesi-ve-hastalarda-dogrulanmasi.

[51] O. Erdoğan *et al.*, "Identification of genomic biomarkers with machine learning for early and differential diagnosis of late-onset Alzheimer's disease (LOAD)," *Alzheimer's Dement.*, vol. 16, no. S2, p. 42558, 2020, doi: 10.1002/alz.042558.

[52] H. Byeon, "Is the random forest algorithm suitable for predicting parkinson's disease with mild cognitive impairment out of parkinson's disease with normal cognition?," *Int. J. Environ. Res. Public Health*, vol. 17, no. 7, 2020, doi: 10.3390/ijerph17072594.

[53] M. Pal and S. Parija, "Prediction of Heart Diseases using Random Forest," *J. Phys. Conf. Ser.*, vol. 1817, no. 1, pp. 0–8, 2021, doi: 10.1088/1742-6596/1817/1/012009.

[54] "SNPnexus," 2022. https://www.snp-nexus.org/v4/.

[55] J. Reimand *et al.*, "Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap," *Nat. Protoc.*, vol. 14, no. 2, pp. 482–517, 2019, doi: 10.1038/s41596-018-0103-9.

[56] "SNiPA," 2022. https://snipa.helmholtz-muenchen.de/snipa3/.

[57] S. Padmanabhan, C. Menni, D. Prabhakaran, and A. F. Dominiczak, "Discovering the genetic determinants of complex diseases," *Curr. Sci.*, vol. 97, no. 3, pp. 385–391, 2009.

[58] C. S. Greene *et al.*, “Understanding multicellular function and disease with human tissue-specific networks,” *HHS Public Access*, vol. 47, no. 6, pp. 569–576, 2016, doi: 10.1038/ng.3259.Understanding.

[59] V. Vasic, K. Barth, and M. H. H. Schmidt, “Neurodegeneration and neuro-regeneration— Alzheimer’s disease and stem cell therapy,” *Int. J. Mol. Sci.*, vol. 20, no. 17, 2019, doi: 10.3390/ijms20174272.

[60] N. A. Shirwany, D. Payette, J. Xie, and Q. Guo, “The amyloid beta ion channel hypothesis of Alzheimer’s disease,” *Neuropsychiatr. Dis. Treat.*, vol. 3, no. 5, pp. 597–612, 2007.

[61] A. Florentinus-Mefailoski, P. Bowden, P. Scheltens, J. Killestein, C. Teunissen, and J. G. Marshall, “The plasma peptides of Alzheimer’s disease,” *Clin. Proteomics*, vol. 18, no. 1, pp. 1–26, 2021, doi: 10.1186/s12014-021-09320-2.

[62] Y. Sun *et al.*, “Role of the Extracellular Matrix in Alzheimer’s Disease,” *Front. Aging Neurosci.*, vol. 13, no. August, pp. 1–11, 2021, doi: 10.3389/fnagi.2021.707466.

[63] F. Sáez-Orellana, J.-N. Octave, and N. Pierrot, “Alzheimer’s Disease, a Lipid Story: Involvement of Peroxisome Proliferator-Activated Receptor α.,” *Cells*, vol. 9, no. 5, May 2020, doi: 10.3390/cells9051215.

[64] Z. Lv, M. Hashemi, S. Banerjee, K. Zagorski, J.-C. Rochet, and Y. L. Lyubchenko, “Assembly of α-synuclein aggregates on phospholipid bilayers.,” *Biochim. Biophys. acta. Proteins proteomics*, vol. 1867, no. 9, pp. 802–812, Sep. 2019, doi: 10.1016/j.bbapap.2019.06.006.

[65] X. Pan, A. C. Kaminga, S. W. Wen, X. Wu, K. Acheampong, and A. Liu, “Dopamine and dopamine receptors in Alzheimer’s disease: A systematic review and network meta-analysis,” *Front. Aging Neurosci.*, vol. 10, no. JUL, pp. 1–14, 2019, doi: 10.3389/fnagi.2019.00175.

[66] Z.-R. Chen, J.-B. Huang, S.-L. Yang, and F.-F. Hong, “Role of Cholinergic Signaling in Alzheimer’s Disease.,” *Molecules*, vol. 27, no. 6, Mar. 2022, doi: 10.3390/molecules27061816.

[67] A. Alhowail, “Molecular insights into the benefits of nicotine on memory and cognition (Review),” *Mol. Med. Rep.*, vol. 23, no. 6, 2021, doi: 10.3892/MMR.2021.12037.

[68] O. Erdoğan, “EnsembleScoringAlgorithm-EnSCAN @ github.com.” 2024, [Online]. Available: https://github.com/erdogan-onur/EnsembleScoringAlgorithm-EnSCAN.

APPENDICES

APPENDIX A

Ensemble Scoring Algorithm Pseudo

Pseudo Algorithm:

```
class SNPDataEnsemble:
    DF_SNVs<-read (SNVs)
    DF_LDs<-read(proxySeach LDs)
    While true:
    initial_input <-Read (Select LD Coefficient)
        if 0 <= initial_input <= 1:
        LDCorrelationCoefficient<-initial_input
        Break
    DF_LDs<-Filter (DF_LDs) by LDCorrelationCoefficient
    DF_SNVs<-Sort (“DIST”)
    LDDict<-findLinkageDisequilibrium (DF_SNVs, DF_LDs)
    DF_SNVs<-assignLinkDiseqLabels (DF_SNVs, LDDict)
    calculateCenteredScore(DF_SNVs)
    calculateFrameFirstScore(DF_SNVs)
    calculateFrameLastScore(DF_SNVs)
```

```
        finalScores (DF_SNVs)
```

```
def findLinkageDisequilibrium (DF_SNVs, DF_LDs):
        Dictionary LDDictList
        For each searched data in DF_SNVs["LOAD RF RF"]:
                If <searched data> is not empty
                        For each reference data in DF_LDs["QRSID"]:
                                If searched data in DF_SNVs=reference data in DF_LDs
                                List RelatedLD<-reference data
                        LDDict <- RelatedLD
```

```
Def assignLinkDiseqLabels (DF_SNVs, LDDict) :
        For each SNV in DF_SNVs:
                LDValues<-Find SNV in LDDict
                If len(LDValues)>0:
                        For each related LD data in LDValues:
                                Append DF_SNVs["Label"]<-LD data
```

```
Def calculateScore(DF_SNVs):
        For each index of searched data in DF_SNVs["LOAD RF RF"]:
                Startframe<-Find startframe while previous rows are not in LD
                Endframe<-Find endframe while next rows are not in LD
                SSforDB <- DF_SNVs[int(startframe): int(endframe) + 1]
                numberofDatabase <- count distinct (SSforDB['DATABASE'])
                preneighbourhood<-Find previous row
```

postneighbourhood<-Find next row

SSforLD <-DF_SNVs[int(preneighbourhood): int(postneighbourhood) + 1]

If searched data is in LD with (preneighbourhood and postneighbourhood)

numberofIndependentLD=1

else if search data is not in LD with (preneighbourhood and postneighbourhood)

numberofIndependentLD=3

else if search data is in LD with (preneighbourhood) but not in LD with (postneighbourhood)

numberofIndependentLD=2

else if search data is not in LD with (preneighbourhood) but in LD with (postneighbourhood)

numberofIndependentLD=2

Score<- (1*numberofDatabase) + (0.01*numberofIndependentLD) + (0.01)

Def maxScore(DF_SNVs):

maxScores=Find max (DF_SNV['Score 1 (up)', 'Score 2(middle)', 'Score 3(bottom)'])

DF_SNVs['Max Scores']<-maxScores

APPENDIX B

Ensemble Model-Based Significant Protein Coding Variants of LOAD

dbSNP	Chromosome	Overlapped Gene	Phenotype
rs2273570	1	CSMD2	Attention-deficit hyperactivity disorder, Brain structure, Alcoholism, Brain, Lipoproteins
rs9425897	1	CSMD2	Attention-deficit hyperactivity disorde, Brain structure, Alcoholism, Brain, Lipoproteins, Alcoholism, Inflammatory Bowel Diseases
rs2821338	1	NR5A2	Pancreatic cancer, Pancreatic Neoplasms, Creatinine, Neuroblastoma, Bone Density, Body Weight
rs10924173	1	KIF26B	Cholesterol, HDL, Stroke, Body Height, Insulin
rs4750306	10	CCDC3	Apolipoproteins C
rs2503848	10	CSGALNACT2	Alzheimer's disease, Monocytes
rs363225	10	SLC18A2	Antidepressant response, Marijuana Abuse|Psychoses, Substance-Induced, Narcolepsy, Bulimia, Tobacco Use Disorder, Alcohol abuse; Nicotine dependence, Parkinson's disease, Bipolar disorder schizophrenia, Alcohol consumption, Financial and psychological risk attitudes,

dbSNP	Chromosome	Overlapped Gene	Phenotype
			Weight Gain, Smoking cessation, Respiratory Function Tests
rs6578869	11	PPFIBP2	Tobacco Use Disorder, Breast cancer
rs4755903	11	TSPAN18	Tobacco Use Disorder, Schizophrenia
rs557098	11	ALDH3B1	-
rs1421555	11	CNTN5	Longevity, Type 2 Diabetes| edema | rosiglitazone, Alzheimer's disease, Psychiatric Disorders,Tobacco Use Disorder, Platelet Aggregation, CD40 Ligand Tunica Media, Heart Rate, Attention Deficit Disorder with Hyperactivity, Stroke, Atrial Fibrillation, Carotid Stenosis, Brain, Frontal Lobe, Body Height, Heart Failure, Hip, Fibrinogen, Blood Coagulation Factors, Myocardial Infarction
rs6490095	12	RNFT2	-
rs4766819	12	FBXW8	Cholesterol, LDL
rs12920363	16	RBFOX1	-
rs4784044	16	NDRG4	Arrhythmias, Cardiac|Death, Sudden, Cardiac|, QT interval, Tobacco Use Disorder
rs2537828	17	PITPNC1	Stroke, Alzheimer Disease
rs9911460	17	NPLOC4	-

dbSNP	Chromosome	Overlapped Gene	Phenotype
rs623755	18	PHLPP1	Narcolepsy, Tobacco Use Disorder
rs16993582	21	RUNX1	Arthritis, Rheumatoid, Leukemia, Myelomonocytic, Chronic, Prostate cancer, Type 2 Diabetes| edema | rosiglitazone, Leukemia, Myeloid, Acute|Myelodysplastic Syndromes|Preleukemia, Tobacco Use Disorder, Hepatopulmonary Syndrome|Liver Cirrhosis, Bilirubin, Stroke, Vitamin D, Esophageal Neoplasms, Phospholipids
rs5753336	22	OSBP2	Macular Degeneration, Type 2 Diabetes| edema | rosiglitazone, Tobacco Use Disorder, Heart Rate
rs2071846	22	CYB5R3	Chronic renal failure|Kidney Failure, Chronic
rs17642119	3	LSAMP	Suicide, Atherosclerosis|Coronary Artery Disease, Tobacco Use Disorder, Mental Competency
rs16863803	3	TPRG1	-
rs4687319	3	FGF12	Tobacco Use Disorder, Serum metabolites, Cleft Lip|Cleft Palate, Tobacco Use Disorder, Uric Acid, Carcinoid Tumor
rs10520536	4	TENM3	-
rs10024098	4	ENPP6	-

dbSNP	Chromosome	Overlapped Gene	Phenotype
rs3816620	5	ADCY2	Chronic renal failure|Kidney Failure, Chronic, Tobacco Use Disorder, Antipsychotic response | Weight Gain, Asthma, Stroke
rs6864906	5	ANKRD33B	-
rs3797390	5	IQGAP2	Longevity, Type 2 Diabetes| edema | rosiglitazone, Tobacco Use Disorder,Smoking cessation, Hearing impairment
rs10515489	5	SPOCK1	Tobacco Use Disorder
rs3776403	5	ARHGAP26	Tobacco Use Disorder, Triglycerides, Diabetes Mellitus, Type 2
rs962459	5	CYFIP2	Asthma|Bronchial Hyperreactivity|Hypersensitivity, Immediate, Heart Rate, Body Height, Heart Rate
rs9366664	6	SYCP2L	Age at menarche/menopause
rs9295676	6	SLC17A2	Gout
rs2268408	6	MDFI	-
rs9296459	6	RUNX2	Osteoporosis, Ossification of Posterior Longitudinal Ligament|Ossification, Heterotopic, Cleidocranial Dysplasia, Cleft Lip|Cleft Palate, Postmenopausal, Tobacco Use Disorder, Bone mineral density, Bone density, Bone size, Alzheimer's disease, Diabetes Mellitus, Type 2, Cleft Lip|Cleft

dbSNP	Chromosome	Overlapped Gene	Phenotype
			Palate|, Height, Insulin, Occipital Lobe, Fibrinogen
rs12197213	6	CRISP1	-
rs2207851	6	PLAGL1	-
rs17300224	7	GNGT1	-
rs17092948	8	DLC1	Type 2 Diabetes| edema | rosiglitazone, Tobacco Use Disorder, Cholesterol, HDL, Bone Density
rs7029570	9	C9orf3	Tobacco Use Disorder, Hemoglobins
rs4978508	9	SNX30	Cleft Lip|Cleft Palate
rs7045548	9	HMCN2	-

APPENDIX C

Ensemble Scoring Algorithm Python Code [68]

```
class SNPDataEnsemble:
    LDCorrelationCoefficient=0.70 #default
    print("Reading SNPs Data (Combined):")
    df_SNPsData = pd.read_excel(io='C:/Users/Onur/Desktop/SNPDataset.xlsx',
                                sheet_name='SNPDataset (Combined) (3)')
    df_SNPsData['ChrNumber'] = df_SNPsData['Chromosome'].str[3:]
    df_SNPsData['Label'] = df_SNPsData['Label'].astype(str)
    # Define the sorter
    sorter = ['chr1', 'chr2', 'chr3', 'chr4', 'chr5', 'chr6', 'chr7', 'chr8', 'chr9', 'chr10', 'chr11',
              'chr12', 'chr13', 'chr14', 'chr15', 'chr16', 'chr17', 'chr18', 'chr19', 'chr20', 'chr21', 'chr22', 'chrX',
              'chrY']
    # Create the dictionary that defines the order for sorting
    sorterIndex = dict(zip(sorter, range(len(sorter))))
    print("Sorter Index Created by Chromosome Number...")
    df_SNPsData['Chr_Rank'] = df_SNPsData['Chromosome'].map(sorterIndex)
    df_SNPsData.sort_values(by=['Chr_Rank','Position'], ascending=[True, True], inplace=False)
    df_SNPsData['AutoNumberAfterSort'] = df_SNPsData.reset_index().index
    print( df_SNPsData.info(verbose=True))
    df_SNPsData = df_SNPsData[
        ["Score UP", "Score Middle", "Score BOTTOM", "Label", "Order", "SNP", "Chromosome","Position", "Chr_Rank",
"DATABASE", "checkPRFRF"]]
    print("Reading SNPs Data (Combined) Completed Successfully...")
    print("--------")
    print("\n Reading LDs Data (Proxy Data:)")
    df_LDsData = pd.read_excel(io='C:/Users/Onur/Desktop/SNPDataset.xlsx', sheet_name='proxySearch.results')
    df_LDsData = df_LDsData[["QRSID", "RSID", "R2", "POS2", "DIST"]]
    print("Reading LDs Data (Proxy Data:) Completed Successfully...")
    print("--------")
    while True:
        initial_input = raw_input(
            " \n (LD correlation Coefficient Identification) Please enter a number between 1 and 0: ").strip()
        if "." in initial_input or initial_input.isnumeric():
            initial_input = float(initial_input)
            if 0 <= initial_input <= 1:
                LDCorrelationCoefficient = initial_input
                break
        else:
            print("Please enter the value correctly, it must be a real number between 0 and 1")
    print("LDs Data is Being Preprocessed Based On Selected Coefficient: " ,LDCorrelationCoefficient)
    a = df_LDsData.sort_values(by=['DIST'], ascending=[False])
    df_LDsData = a[(a['QRSID'] != a['RSID']) & (a['R2'] > LDCorrelationCoefficient) ]



    def writeToExcelFile(x, processType):
        if processType == 'SNP':
            with ExcelWriter('path_to_file_SNP.xlsx') as writer:
                x.to_excel(writer)
                #print('SNPs Succesfully Exported To Excel')
        if processType == 'LD':
            #string_value = str(LD)
            #string_value=string_value.replace(".", "_")
            filename=('path_to_file_LD_Reference_Table.xlsx')
            with ExcelWriter(filename) as writer:
                x.to_excel(writer)
                print('Based on User Input, LDs Succesfully Exported To Excel File')
        if processType == 'LD-Relationship':
            with ExcelWriter('path_to_file_SNP-LD_Relation.xlsx') as writer:
                x.to_excel(writer)
                print('SNPs are Succesfully Mapped by LDs and Exported To Excel')
```

```
    def findLinkageDisequilibrium (df_SNPsData, df_LDsData):
        LDDict={}
        for index, row in df_SNPsData.iterrows():
            RelatedLD=[]
            if not pd.isna(row ["checkPRFRF"]):
                print("Processing for # of ", index, "->", row["checkPRFRF"])
                for indexLD,rowLD in df_LDsData.iterrows():
                    if (row ["checkPRFRF"]==rowLD["QRSID"]):
                        RelatedLD.append(rowLD['RSID'])
                print (RelatedLD)
                LDDict[row ["checkPRFRF"]]=RelatedLD
        w = csv.writer(open("DeterminedLDOutput.csv", "w"))
        for key, val in LDDict.items():
            w.writerow([key, val])
        return LDDict



    def calculateCenteredScore(SNPsData):
        for index, row in SNPsData.iterrows():
            if not pd.isna(row["checkPRFRF"]):
                if index>0:
                    prev=index-1
                    while ((str(index) + ".LD") in SNPsData.at [prev,'Label']):
                        prev = prev - 1
                    startframe=prev
                else:
                    startframe=prev
                if index<len(SNPsData)-1:
                    nxt = index + 1
                    while ((str(index) + ".LD") in SNPsData.at[nxt, 'Label']):
                        nxt = nxt + 1
                    endframe = nxt
                else:
                    endframe=None
                file1 = open("MiddleScoreStartEnd.txt", "a")
                st = ("For #" + str(index) + " index value [Start: " + str(startframe) + ", End:" + str(endframe) +
"]\n")
                file1.writelines (st)
                file1.close()
                if(isinstance(startframe, int) & isinstance(endframe, int)):
                    SSforDB = SNPsData[int(startframe): int(endframe) + 1]
                    numberofDatabase = SSforDB['DATABASE'].nunique()
                if index>0:
                    preneighbourhood= index-1
                else: preneighbourhood=None
                if index<len(SNPsData)-1:
                    postneighbourhood=index+1
                else:
                    postneighbourhood=None
                if (isinstance(preneighbourhood, int) & isinstance(postneighbourhood, int)):
                    SSforLD = SNPsData[int(preneighbourhood): int(postneighbourhood) + 1]
                    if SSforLD.at[index, 'Label'] != 'nan':
                        if ((str(index) + ".LD") in SNPsData.at[preneighbourhood, 'Label']) and ((str(index) + ".LD") in
SNPsData.at[postneighbourhood, 'Label']):
                            numberofIndependentLD=1
                        elif ((str(index) + ".LD") not in SNPsData.at[preneighbourhood, 'Label']) and ((str(index) +
".LD") not in SNPsData.at[postneighbourhood, 'Label']):
                            numberofIndependentLD=3
                        elif ((str(index) + ".LD") in SNPsData.at[preneighbourhood, 'Label']) and ((str(index) + ".LD")
not in SNPsData.at[postneighbourhood, 'Label']):
                            numberofIndependentLD=2
                        elif ((str(index) + ".LD") not in SNPsData.at[preneighbourhood, 'Label']) and ((str(index) +
".LD") in SNPsData.at[postneighbourhood, 'Label']):
                            numberofIndependentLD=2
                    elif SSforLD.at[index, 'Label'] == 'nan':
                        if (SSforLD.at[index, 'Label'] ==SSforLD.at[preneighbourhood, 'Label']) and (SSforLD.at[index,
'Label'] ==SSforLD.at[postneighbourhood, 'Label']):
                            numberofIndependentLD=3
                        elif (SSforLD.at[index, 'Label'] !=SSforLD.at[preneighbourhood, 'Label']):
                            numberofIndependentLD=3
                        elif (SSforLD.at[index, 'Label'] !=SSforLD.at[postneighbourhood, 'Label']):
                            numberofIndependentLD=3
```

```
                    else: numberofIndependentLD="*"
                    if (isinstance(preneighbourhood, int) & isinstance(postneighbourhood, int) & isinstance(startframe,
int) & isinstance(endframe, int)):
                        SNPsData.at[index, 'Score Middle'] = (1*numberofDatabase) + (1/10*numberofIndependentLD) +
(1/100)
                    else:
                        SNPsData.at[index, 'Score Middle'] =None



    def calculateFrameFirstScore (SNPsData):
        for index, row in SNPsData.iterrows():
            if not pd.isna(row["checkPRFRF"]):
                if index==0:
                    startframe=index
                elif index>0:
                    prev=index-1
                    counter=0
                    while ((str(index) + ".LD") in SNPsData.at[prev, 'Label']):
                        counter = counter + 1
                        prev=prev-1
                        if prev==0:
                            break
                    if counter > 0:
                        startframe = prev+1
                    else:
                        startframe = index
                if index<len(SNPsData)-2:
                    nxt = index + 1
                    counter=0
                    while ((str(index) + ".LD") in SNPsData.at[nxt, 'Label']):
                        counter=counter+1
                        nxt = nxt + 1
                        if nxt>len(SNPsData)-2:
                            break
                    if counter==0 and index<len(SNPsData)-2 :
                        endframe = index + 2
                    elif    counter==0 and index>=len(SNPsData)-2:
                        endframe=None
                    elif counter>0 and nxt>len(SNPsData)-2:
                        endframe=None
                    else:
                        endframe=nxt+1
                else:
                    endframe=None
                file2 = open("FrameFirstStartEnd.txt", "a")
                st = ("For #" + str(index) + " index value [Start: " + str(startframe) + ", End:" + str(endframe) +
"]\n")
                file2.writelines (st)
                file2.close()
                if (isinstance(startframe, int) & isinstance(endframe, int)):
                    SSforDB = SNPsData[int(startframe): int(endframe) + 1]
                    numberofDatabase = SSforDB['DATABASE'].nunique()
                if index>0:
                    preneighbourhood= index-1
                else: preneighbourhood=None
                if index<len(SNPsData)-1:
                    postneighbourhood=index+1
                else:
                    postneighbourhood=None
                if (isinstance(preneighbourhood, int) & isinstance(postneighbourhood, int)):
                    SSforLD = SNPsData[int(preneighbourhood): int(postneighbourhood) + 1]
                    if SSforLD.at[index, 'Label'] != 'nan':
                        if ((str(index) + ".LD") in SNPsData.at[preneighbourhood, 'Label']) and ((str(index) + ".LD") in
SNPsData.at[postneighbourhood, 'Label']):
                            numberofIndependentLD=1
                        elif ((str(index) + ".LD") not in SNPsData.at[preneighbourhood, 'Label']) and ((str(index) +
".LD") not in SNPsData.at[postneighbourhood, 'Label']):
                            numberofIndependentLD=3
                        elif ((str(index) + ".LD") in SNPsData.at[preneighbourhood, 'Label']) and ((str(index) + ".LD")
not in SNPsData.at[postneighbourhood, 'Label']):
                            numberofIndependentLD=2
                        elif ((str(index) + ".LD") not in SNPsData.at[preneighbourhood, 'Label']) and ((str(index) +
".LD") in SNPsData.at[postneighbourhood, 'Label']):
                            numberofIndependentLD=2
                    elif SSforLD.at[index, 'Label'] == 'nan':
```

```
                        if (SSforLD.at[index, 'Label'] ==SSforLD.at[preneighbourhood, 'Label']) and (SSforLD.at[index,
'Label'] ==SSforLD.at[postneighbourhood, 'Label']):
                            numberofIndependentLD=3
                        elif (SSforLD.at[index, 'Label'] !=SSforLD.at[preneighbourhood, 'Label']):
                            numberofIndependentLD=3
                        elif (SSforLD.at[index, 'Label'] !=SSforLD.at[postneighbourhood, 'Label']):
                            numberofIndependentLD=3
                    else: numberofIndependentLD="*"
                    if (isinstance(preneighbourhood, int) & isinstance(postneighbourhood, int) & isinstance(startframe,
int) & isinstance(endframe, int)):
                        SNPsData.at[index, 'Score UP'] = (1*numberofDatabase) + (1/10*numberofIndependentLD) + (1/100)
                    else:
                        SNPsData.at[index, 'Score UP'] =None



    def calculateFrameLastScore(SNPsData):
        for index, row in SNPsData.iterrows():
            if not pd.isna(row["checkPRFRF"]):
                if index<2:
                    startframe=None
                elif index>=2:
                    prev=index-1
                    counter=0
                    while ((str(index) + ".LD") in SNPsData.at[prev, 'Label']):
                        counter = counter + 1
                        prev=prev-1
                        if prev==0:
                            break
                    if counter == 0:
                        startframe = index - 2
                    if counter > 0:
                        if prev + 1 > 1:
                            startframe = prev-1
                        else:
                            startframe = None
                else:
                    startframe=None
                if index == len(SNPsData) - 1:
                    endframe=index
                elif index>1 and index < len(SNPsData) - 1 :
                    nxt = index + 1
                    counter = 0
                    while ((str(index) + ".LD") in SNPsData.at[nxt, 'Label']):
                        counter = counter + 1
                        nxt = nxt + 1
                        if nxt == len(SNPsData) - 1:
                            break
                    if counter > 0:
                        endframe = nxt - 1
                    elif counter==0:
                        endframe = index
                else:
                    endframe=None
                file3 = open("FrameLastStartEnd.txt", "a")
                st = ("For #" + str(index) + " index value [Start: " + str(startframe) + ", End:" + str(endframe) +
"]\n")
                file3.writelines (st)
                file3.close()
                if (isinstance(startframe, int) & isinstance(endframe, int)):
                    SSforDB = SNPsData[int(startframe): int(endframe) + 1]
                    numberofDatabase = SSforDB['DATABASE'].nunique()
                if index>0:
                    preneighbourhood= index-1
                else: preneighbourhood=None
                if index<len(SNPsData)-1:
                    postneighbourhood=index+1
                else:
                    postneighbourhood=None
                if (isinstance(preneighbourhood, int) & isinstance(postneighbourhood, int)):
                    SSforLD = SNPsData[int(preneighbourhood): int(postneighbourhood) + 1]
                    if (isinstance(preneighbourhood, int) & isinstance(postneighbourhood, int)):
                        SSforLD = SNPsData[int(preneighbourhood): int(postneighbourhood) + 1]
                        if SSforLD.at[index, 'Label'] != 'nan':
                            if ((str(index) + ".LD") in SNPsData.at[preneighbourhood, 'Label']) and (
                                    (str(index) + ".LD") in SNPsData.at[postneighbourhood, 'Label']):
```

```
                            numberofIndependentLD = 1
                        elif ((str(index) + ".LD") not in SNPsData.at[preneighbourhood, 'Label']) and (
                                (str(index) + ".LD") not in SNPsData.at[postneighbourhood, 'Label']):
                            numberofIndependentLD = 3
                        elif ((str(index) + ".LD") in SNPsData.at[preneighbourhood, 'Label']) and (
                                (str(index) + ".LD") not in SNPsData.at[postneighbourhood, 'Label']):
                            numberofIndependentLD = 2
                        elif ((str(index) + ".LD") not in SNPsData.at[preneighbourhood, 'Label']) and (
                                (str(index) + ".LD") in SNPsData.at[postneighbourhood, 'Label']):
                            numberofIndependentLD = 2
                    elif SSforLD.at[index, 'Label'] == 'nan':
                        if (SSforLD.at[index, 'Label'] == SSforLD.at[preneighbourhood, 'Label']) and (
                                SSforLD.at[index, 'Label'] == SSforLD.at[postneighbourhood, 'Label']):
                            numberofIndependentLD = 3
                        elif (SSforLD.at[index, 'Label'] != SSforLD.at[preneighbourhood, 'Label']):
                            numberofIndependentLD = 3
                        elif (SSforLD.at[index, 'Label'] != SSforLD.at[postneighbourhood, 'Label']):
                            numberofIndependentLD = 3
                    else:
                        numberofIndependentLD = "*"
                if (isinstance(preneighbourhood, int) & isinstance(postneighbourhood, int) & isinstance(startframe,
int) & isinstance(endframe, int)):
                    SNPsData.at[index, 'Score BOTTOM'] = (1*numberofDatabase) + (1/10*numberofIndependentLD) +
(1/100)
                else:
                    SNPsData.at[index, 'Score BOTTOM'] = None



    def maxScore(df_SNPsData):
        maxScores=df_SNPsData[['Score UP', 'Score Middle', 'Score BOTTOM', 'SNP']].max(axis=1)
        df_SNPsData['Max Scores']=maxScores
        return df_SNPsData



    def assignLinkDiseqLabels (SNPsData, LDDict):
        for index, row in SNPsData.iterrows():
            if not pd.isna(row["checkPRFRF"]):
                key=row["checkPRFRF"]
                LDValues=LDDict[row["checkPRFRF"]]
                print (key, ": ",  LDValues, "->", len (LDValues) )
                print('\n')
                if len(LDValues)>0:
                    for i in LDValues:
                        ind=int(SNPsData[SNPsData["SNP"] == i].index.values)
                        if SNPsData.at [ind,'Label']=='nan':
                            st = str(index) + ".LD"
                            SNPsData.at[ind, 'Label'] = st
                            SNPsData.at[index, 'Label'] = st
                        else:
                            st = (SNPsData.at [ind,'Label']) + ", " +  str(index) + ".LD"
                            SNPsData.at[ind, 'Label'] = st
                            SNPsData.at[index, 'Label'] = st
        return SNPsData



    if __name__ == "__main__":
        print('\n',"Scoring Has Been Started.....")
        print("*****************************")
        print("After filtering is implemented over LD Pairwise Data: ", len(df_LDsData),'\n')
        print("After combining Significant SNPs from Different Databases: ",len(df_SNPsData),'\n')
        print ("RF importances have been validated by a Second-run RF: ", df_SNPsData.checkPRFRF.notnull().sum(), '\n' )
        writeToExcelFile(df_SNPsData, "SNP")
        writeToExcelFile(df_LDsData, "LD")
        LDDictionary=findLinkageDisequilibrium (df_SNPsData,df_LDsData)
        print(LDDictionary)
        df_SNPsData=assignLinkDiseqLabels(df_SNPsData, LDDictionary)
        writeToExcelFile(df_SNPsData, "LD-Relationship")
        calculateCenteredScore(df_SNPsData)
        writeToExcelFile(df_SNPsData, "LD-Relationship")
        calculateFrameFirstScore(df_SNPsData)
        writeToExcelFile(df_SNPsData, "LD-Relationship")
        calculateFrameLastScore(df_SNPsData)
```

```
        writeToExcelFile(df_SNPsData, "LD-Relationship")
        finalScores=maxScore(df_SNPsData)
        writeToExcelFile(df_SNPsData, "LD-Relationship")
        temp = df_SNPsData.groupby(['Max Scores'])['Max Scores'].count().to_frame('Count').reset_index()
        formattedPoints = ["%.2f" % member for member in temp.iloc[:, 0]]
        points = list(formattedPoints)
        print (points)
        count = list(temp.iloc[:, 1])
        print (count)
        temp['Max Scores'] = 'p' + temp['Max Scores'].astype(str)
        bars=plot.bar(points, count, color ='maroon', width = 0.3, align='center')
        plot.xlabel("Available Scores")
        plot.ylabel("Observation")
        plot.title("Distribution of Points")
        for bar in bars:
            yval = bar.get_height()
            plot.text(bar.get_x(), yval + .005, yval)
        filename = 'ScoreDistribution - ' + 'LDCoef' + str(initial_input) + '.png'
        plot.savefig(filename, dpi=200,bbox_inches='tight')
        plot.show()
```

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Erdoğan, Onur

Nationality: Turkish (TC)

EDUCATION

Degree	Institution	Year of Graduation
MS	METU, Medical Informatics	2012
BS	Hacettepe University, Statistics	2008
High School	75. Yıl Lisesi (Yabancı Dil Ağırlıklı)	2003

WORK EXPERIENCE

Year	Place	Enrollment
2022-Present	HAVELSAN	Team Manager
2019-2022	TÜBİTAK	Head of Quality and Strategy Management Department
2011-2019	TÜBİTAK	Measurement, Analysis, Business Intelligence Specialist
2010 May	YARGITAY	Programmer
2008 October	SAD MARKETING	Area Manager

FOREIGN LANGUAGES

Native Turkish, Advanced English

SKILLS

- **Statistical Tools**
    - SPSS, Minitab, R

- **Programming Languages**
    - Python, Java, Turbo Pascal, Visual Basic 6.0/.Net, C, Matlab, PHP
- Software Development Processes
    - SDLC, SAFe, Scrum, Kanban, CMMI DEV
- **Database, Data Warehouse, and Business Intelligence**
    - MS SQL, MySQL, MS BI (Analysis Services, Integration Services), Qlik Sense, Tableau
- **Management Tools**
    - Atlassian Suite (Jira, Confluence, etc.), SVN, IBM Jazz, Github
- **Networking**
    - FTP Server, DNS Server, Active Directory

PUBLICATIONS, PRESENTATIONS, POSTERS, PANELS

- Yaldız, B., **Erdoğan, O.**, Rafatov, S., Iyigün, C., & Aydın Son, Y. 2024. "Revealing Third-Order Interactions Through the Integration of Machine Learning and Entropy Methods in Genomic Studies." *BioData Mining*, 17(1), 1-17. https://doi.org/10.1186/s13040-024-00355-3.

- **Erdoğan, O.** 2012. "Yazılım Geliştirme Süreçlerinin İyileştirilmesinde İstatistiksel Süreç Kontrolü ve Kök Neden Saptama." Uzmanlık Tezi, *TÜBİTAK*.
- **Erdoğan, O.**, Esme, M., Balci, C., et al. 2020. "Identification of Genomic Biomarkers with Machine Learning for Early and Differential Diagnosis of Late-Onset Alzheimer's Disease (LOAD)." *Alzheimer's Dementia*, 16(S2), 42558. https://doi.org/10.1002/alz.042558.

- **Erdoğan, O.**, Pekkaya, E., & Gök, H. 2017. "More Effective Sprint Retrospective with Statistical Analysis." *Journal of Software: Evolution and Process*. Published online. https://doi.org/10.1002/smr.1933.

- Şimşek, H., **Erdoğan, O.**, Çınar Özen, N., & Tekmen, F. 2017. "An Industrial Experience in CMMI L5 Improvement Opportunity: Implementing Usability." In *European Organization for Quality (EOQ) 2017*, 8.

- **Erdoğan, O.**, Ünlü, N., Dikici, A., & Gök, H. 2017. "Yazılımda Yüksek Olgunluk Paneli, Panelist." *TÜBİTAK BİLGEM YTE tarafından BTK Konferans Salonu'nda düzenlenen özel sektör ve kamu kuruluşu profesyonellerinin iştirak ettiği Yazılımda Yüksek Olgunluk Etkinliği*.

- **Erdoğan, O.** 2012. “CMMI-DEV Yüksek Olgunluk Yaklaşımında Kalite ve Süreç Performans Hedeflerinin İzlenebilirliği ve Süreç İyileştirme Faaliyetleri.” Başuzmanlık Tezi, *TÜBİTAK*.

- **Erdoğan, O.**, & Aydın Son, Y. 2014. "Predicting the Disease of Alzheimer with SNP Biomarkers and Clinical Data Using Data Mining Classification Approach: Decision Tree." In *Studies in Health Technology and Informatics*, vol. 205, 511-515. IOS Press BV. https://doi.org/10.3233/978-1-61499-432-9-511.

- **Erdoğan, O.**, & Aydın Son, Y. 2013. "Bioinformatics of Disease and Treatment: A Decision Support Model Based on Integration of Genomic and Clinical Findings for the Differential Diagnosis of Alzheimer’s Disease." In *21st Annual International Conference on Intelligent Systems for Molecular Biology (ISMB)*. Retrieved from http://www.iscb.org/cms_addon/conferences/ismbeccb2013/posterlist.php?cat=A.

- Günel, B. D., **Erdoğan, O.**, Dikici, A., Çağlayan, K., & Güngör, M. K. 2013. "Learner Control Charts: A Structured Way for Finding Common Causes with Decision Trees." In *EuroSPI*, 11.

- **Erdoğan, O.** 2012. "Predicting the Disease Alzheimer (AD) With SNP Biomarkers and Clinical Data-Based Decision Support System Using Data Mining Classification Approaches." Retrieved from http://etd.lib.metu.edu.tr/upload/12614832/index.pdf.

- Günel, B. D., **Erdoğan, O.**, Doğan, G. H., et al. 2012. "Sistem Testlerinde İstatistiksel Süreç Kontrolünün Kullanımı." In *Ulusal Yazılım Mühendisliği Sempozyumu (UYMS)*, 48-53.

AWARDS

- Special Mention Poster Award 2017

  Received at the Institute Day and Open Research Day for a research poster presented at the Informatics Institute.

- TÜBİTAK ARDEB 1003 Program (Grant No: SBAG-216S468)

  Contributed as a scholarship holder to a TÜBİTAK 1003 funded research project based on my doctoral studies.

- TÜBİTAK Institutional Development Award 2014

  Recognized for contributions to TÜBİTAK BİLGEM YTE’s achievement of a high maturity level in CMMI.