Instituto Tecnológico y de Estudios Superiores de Monterrey

Monterrey Campus

Escuela de Ingeniería y Ciencias

Identifying models of DNA polymorphisms associated with Alzheimer's Disease using Step-Wise and Genetic Algorithms from GWAS data

A thesis presented by

Brissa Lizbeth Romero Rosales

Submitted to the
Escuela de Ingeniería y Ciencias
in partial fulfillment of the requirements for the degree of

Master of Science

in

Computer Science

Monterrey, Nuevo León, May, 2019

Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Monterrey

The committee members. hereby, certify that have read the thesis presented by Brissa Lizbeth Romero Rosales and that it is fully adequate in scope and quality as a partial requirement for the degree of Master of Science in Computer Sciences.

Dr. Víctor Manuel Treviño Alvarado
Instituto Tecnológico y de Estudios Superiores de Monterrey
Principal Advisor

Dr. Edgar Emmanuel Vallejo Clemente
Instituto Tecnológico y de Estudios Superiores de Monterrey
Committee Member

Dra. María Guadalupe Moreno Treviño
Universidad de Monterrey
Committee Member

Dr. Rubén Morales Menéndez
Associate Dean of Graduate Studies
Escuela de Ingeniería y Ciencias

Monterrey, Nuevo León, May, 2019

Declaration of Authorship

I, Brissa Lizbeth Romero Rosales, declare that this thesis titled, Identifying models of DNA polymorphisms associated with Alzheimer's Disease using Step-Wise and Genetic Algorithms from GWAS data and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Brissa Lizbeth Romero Rosales
Monterrey, Nuevo León, May, 2019

Dedication

To those people who have suffered from the sad and devastating Alzheimer's disease and for those who have felt the pain of seeing a loved one deteriorate at the expense of this awful disorder but who have dedicated their care, love, and patience. Hopefully one day this disease will cease to be incurable.

Acknowledgements

First of all, I would like to thank my mom Lidia for giving me her unconditional support, for encouraging me to fulfill my goals and dreams, for being the best mom, I would never finish thanking her for everything she has done for me. I would also like to thank my dad Luis, for encouraging me despite the distance and for supporting my decisions. I would also like to thank my sister Yoselin and my brother Jair, because they are always present despite the distance, for their love and for believing in me, for being a motivation to continue every day.

I also want to thank the love of my life, Ricardo, for believing in me, for encouraging me to continue fulfilling goals, for his patience when I found myself in times of stress and for showing me his love every day.

A special mention for my advisor Dr. Victor Treviño for being a fundamental piece in this investigation, thank you for devoting your time and for being patient with me. I also want to thank the members of my committee, Dr. Edgar Vallejo and Dr. Maria Moreno, for their time and advice. I would also like to express my gratitude to Dr. José Tamez who, despite not participating directly in this research, advised me and supported me with the management of FRESA.CAD.

Finally, I would like to thank Tecnológico de Monterrey for its support with a 100% scholarship for my master's degree studies, and CONACyT for support for living expenses during these months, without this help none of this would be possible.

Identifying models of DNA polymorphisms associated with Alzheimer's Disease using Step-Wise and Genetic Algorithms from GWAS data

by

Brissa Lizbeth Romero Rosales

Abstract

Alzheimer's disease is a neurodegenerative disorder that involves cognitive deterioration accompanied by memory loss and inability to reason, affecting the patient's ability to carry out daily activities. This disorder is caused by genetic, environmental and lifestyle factors. The determination of the genetic factors is very important because the disease can be prognosticated and therefore treated before it appears. However, despite research efforts and many putative detections using univariate analyses, only the APOE gene has been plentiful validated as a risk factor associated with late-onset Alzheimer's disease. Thus, the problem of missing heritability arises, implying that only one gene does not determine the heritability of a disorder, but the combined effect of genes could better explain it. Genome-Wide Association Studies (GWAS) traditionally use univariate techniques to determine the association between markers and diseases. This research proposes the use of machine learning techniques based on GWAS data to identify sets of polymorphisms that maximize discrimination between cases and controls. This document explains the traditional strategies and theoretical bases that support this research. It presents previous works that apply multivariate methods for the prediction of different diseases and treatments, and their most representative characteristics are considered the basis to inspire a new solution.

The proposed methodology includes obtaining genetic data and a pre-processing stage. Afterward, the process involves several quality control procedures that filter samples and SNPs to reduce the number of false positives and false negatives. Next, a chi-squared association test with kinship correction is performed to pre-select markers. Predictive models are built using wrapper and embedded computational methods. The first wrapper method used is BSWiMS, which is based on statistics and procedures of forward and backward selection to generate a logistic model. Its best AUC was 0.689. The second wrapper method used is based on stochastic search and was an ensemble of Genetic Algorithms coupled to a Support Vector Machine classifier followed by a Forward Selection that achieved a maximum AUC of 0.716. The third algorithm used is LASSO, one of the most well-known embedded methods, which use L1-regularization and performs a feature selection process in the training stage of the model. This classifier achieved an AUC of 0.8005. This study incorporates the analysis of poorly classified samples in predictive models as a strategy to build higher predictive models. The best result obtained with the mixed model of the variants of previous models outperformed the others with an AUC of 0.842. This result is promising since the model generated with LASSO showed the highest discrimination between classes, based solely on genetic data. The biological relevance of the markers of the models is presented through their association with their respective gene. The models replicated variants previously associated with Alzheimer's disease, especially on chromosome 19 close to the APOE gene.

List of Figures

List of Tables

Contents

1 Introduction

1.1 Brief Overview

Dementia is a set of symptoms associated with cognitive deterioration. It affects memory, thinking, behaviour and in general the ability to perform daily activities. According to the World Health Organization, around 50 million people have dementia and there are about 10 million new cases every year. Dementia affects both patients and their families but also has a social and economic impact. In 2005, the total global societal cost of dementia was estimated in USD $818 billion.

Alzheimer's Disease (AD) is the most common form of dementia in the elderly accounting for 60-70% of dementia cases [3]. The disease has two subtypes based on the age of onset: early-onset AD (EOAD) and late-onset AD (LOAD), details are shown in table 1.1. EOAD is present in 5% of cases [6], with an age onset from 30's to mid-60s and the genes associated are presenilin 1 (PSEN1), presenilin 2 (PSEN2) and amyloid precursor protein (APP) [23]. On the other hand, LOAD occurs after mid-60s and has a 90-95% of incidence [6]. The gene confirmed as a risk factor for LOAD is apolipoprotein E (APOE $\varepsilon 4$) [28].

	Early-Onset AD (Complex Inheritance)	**Late-Onset AD (Complex Inheritance)**
Cause	Genetic and environmental risk factors	Genetic and environmental risk factors
Genes	APP, PSEN1, PSEN2	APOE $\varepsilon 4$
Age at Onset	<65 years	>65 years
Proportion of cases	$\sim$5%	$\sim$90-95%

Table 1.1: Genetics and Alzheimer's Disease. Information synthesized from Bekris, L. et al. (2010) Genetics of Alzheimer Disease [28] and Turan, A. (2010) Late onset Alzheimer's disease in older people [6].

LOAD is a complex disorder that is caused for both genetic and environmental factors [2], as shown in figure 1.1. Heritability, that is, the fraction of patients that can be explained by known genes, of this disease is around 58-79% [31]. Since AD has no cure [28], understanding the genes involved in the evolution of the disease will serve as a guide to identify subjects at risk for early treatment and prevention.

Genome-Wide Association Studies (GWAS) face this problem by measuring and analyzing DNA sequence variations across the human genome [55]. The objective of this method

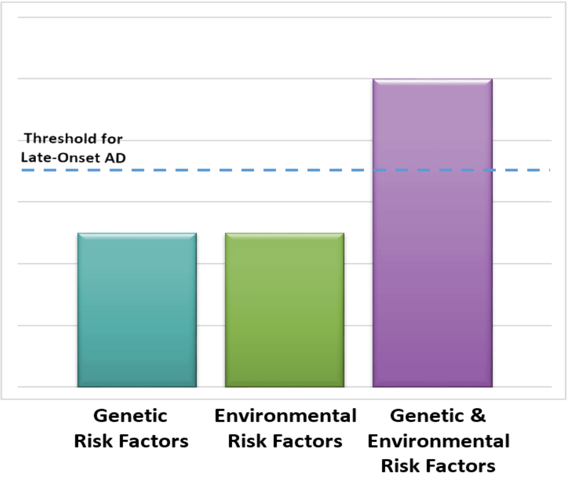


Figure 1.1: Late-onset AD caused by a combination of genetic and environmental factors. Taken from National Centralized Repository For Alzheimer's Disease and Related Dementias (2015) The Genetics of Alzheimer's Disease [2].

is to identify association between Single Nucleotide Polymorphisms (SNPs) and a phenotype in a population (AD in this case). A SNP is a single base-pair change in the DNA sequence that occurs in more than 1% of a population [51]. The most common approach of a GWAS is based on a case-control setup, where the SNPs frequencies across cases and controls are compared to identify those markers with higher frequency in cases than in controls. The association analysis is performed by Chi-squared test or logistic regression assuming independence between variants. However, the genetic markers identified by GWAS only explain a small proportion of the heritability of complex diseases and they do not take into account the interactions between variants [58]. Thus, a multivariate approach is suggested to consider interactions and combinations between SNPs and agreggate their small effects to achieve higher predictive power [46].

1.2 Problem Definition

In medicine, biology, and genetics, there is an urgent need to understand the relationship between genetic markers and phenotypes. The objective of finding variants associated with complex diseases is to improve the quality of life of the affected people through the generation of specialized therapies or medications, as well as to seek to prevent the disease with an early diagnosis. Thanks to advances in genomic technology, such as those based on GWAS, it is relatively easy to obtain genetic data to perform in-depth analysis. However, the problem of having a large number of predictors and few samples remain. Traditional methods for analyzing GWAS data, such as Chi-squared test and logistic regression are based on statistical techniques that evaluate variables one by one for their ability to discriminate between groups of samples (i.e. cases vs. controls) [9]. Although these univariate methods have been successful in identifying hundreds of genomic regions, a large fraction of their heritability

Genotype	**AA**	**AB**	**BB**	**Total**
Case	r0	r1	r2	R
Control	s0	s1	s2	S
Total	n0	n1	n2	N

Table 1.2: Observed genotype counts for cases and controls.

remains unexplained. Possible explanations include that interactions between markers when predicting a disease are not considered [12]. AD is a complex disease, so its prediction could be influenced by multiple genetic variants.

In general, neurodegenerative disorders are a challenge in terms of prediction because pathological information cannot be easily accessed [34] (the only way to have a definitive diagnosis is through an autopsy), to face this issue, practical and non-invasive solutions must be developed to help in the diagnosis.

This section will briefly present the traditional approaches used to identify markers by univariate methods, as well as the strengths and weaknesses of each method when facing the problem posed.

1.2.1 Chi-squared Test

A chi-squared test is a traditional approach used to determine if there is a significant association between categorical variables. When this method is applied to GWAS data, it measures the deviation from independence expected under the null hypothesis that there is no association between genotypes and phenotypes [55]. In a case/control setup, we have a dichotomous variable indicating the phenotype of the subject (0 for control, 1 for case). As well, there are three genotypes indicating the number of alleles in a genetic position. One allele is inherited from each parent. If the genotype has two identical dominant alleles it is called homozygous dominant and is represented as AA or 0. Conversely, the homozygous recessive carries two copies of the same recessive allele and can be identified as BB or 2. On the other hand, if the genotype has two different alleles it is known as heterozygous and is denoted as AB or 1.

The genotype-phenotype information can be summarized in a contingency table of size $(d+1) \times 3$, where d represents the number of diseases (the traditional approach analyzes only one disease, $d = 1$) [60]. An example of the contingency table is shown in table 1.2.

The statistic test is a chi-square random variable χ^2 defined by the following equation:

$$\chi^2 = \sum(\frac{(O - E)^2}{E}) \tag{1.1}$$

Where O is the observed cell count and E is the expected value. Under independence, χ^2 is distributed as chi-squared with 2 degrees of freedom, where $df = (rows - 1) * (columns - 1)$.

A p-value indicates the significance of the difference in frequency of the allele tested between cases and controls. The smaller the p-value, the stronger the evidence against the null hypothesis. Results are often displayed in a Manhattan plot with $-log10(p - value)$

plotted against the position in the genome. The standard process is summarized in figure 1.2.

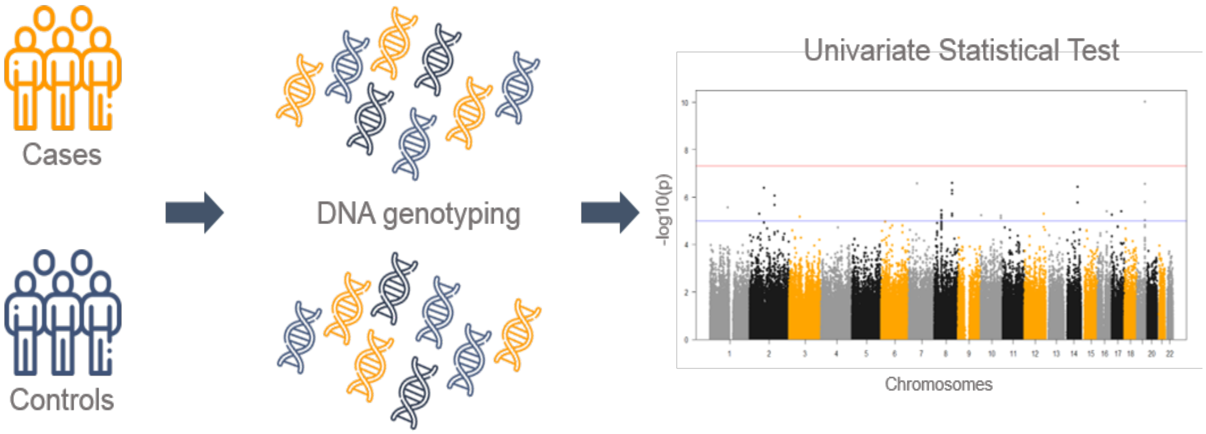


Figure 1.2: Typical methodology to identify variants with higher frequency in cases than controls. P-values are assigned to each marker and represented in a Manhattan plot.

An advantage of a chi-squared test is its robustness and its computational efficiency. However, it does not capture small interactions between markers, which can be informative to a disease. Another disadvantage is that several studies suggest a lack of reproducibility in univariate methods [35].

1.2.2 Logistic Regression

Another common univariate approach used in association studies is logistic regression. An advantage of this method is it's flexibility, because it allows for adjustment through covariates (i.e. gender, age and height) [40] and can calculate adjusted odds ratios, which is a measure of effect size [55]. Logistic regression models the probability that the outcome (phenotype) belongs to a particular predictor (marker) [19]; it can be written as $Pr(Y = 1|x)$, where Y is a categorical variable indicating the phenotype and x is a genotype. The model is constructed with the following equation:

$$log\left(\frac{P(Y=1|X_1,X_2)}{1-P(Y=1|X_1,X_2)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{1.2}$$

In the model, two dummy variables are used to represent genotypes. The first one is denoted by X_1 and corresponds to heterozygous and the second one is X_2 which is used for the homozygous. Details are shown in table 1.3.

The parameters β_0, β_1 and β_2 are estimated through maximum likelihood, which is an iterative method that maximizes the function [24]. Aftterwards a hypothesis test is performed to evaluate the fit of each model and return a p-value for each marker.

Genotype	X_1	X_2
AA	0	0
AB	1	0
BB	0	1

Table 1.3: Dummy variables representation for logistic regression.

1.2.3 Improvement over univariate approaches

As can be seen in the overview and problem description sections, there is huge need to analyze the markers in a different way than univariate methods which only consider a few loci affecting the trait of interest [44]. Although some AD risk genes have been identified, they have not been sufficient to have a better understanding of such a complicated disease [27]. For this reason, the development of new technologies in the area of bioinformatics is required, to look at the problem from another perspective where association analysis can be carried out taking into account the small effects of each variant that maximizes the prediction.

A possible improvement could be made if the advantages of the univariate methods are taken and coupled to multivariate strategies. The problem of having hundreds of thousands of markers and only thousands of samples remains a limitation. Therefore, the reduction of dimensionality could be carried out through statistical analysis, which has shown precision and efficient computation times. Once the number of variants has been reduced, multivariate methods could be applied to explore the combinations that maximize the prediction of the disease.

1.3 Motivation

As mentioned in previous paragraphs, the development of genetic technology spreads fast. As a consequence, clinical and biological data is becoming more accessible, but data needs to be analyzed in different ways to exploit their embed knowledge for human benefits. Therefore, more powerful techniques must be used to analyze, interpret, and transform these data into useful information [61]. Machine Learning (ML) methods have demonstrated to be convincing when faced with a large number of attributes and a few samples. ML is a sub-area of artificial intelligence that incorporates computer science and statistical techniques, where the objective is to build systems that improve automatically through experience [29].

ML techniques have been applied in biomedicine, genetics, image analysis and more recently in genotypic data from GWAS. In general, ML algorithms seek to generate combinations of attributes to predict outcomes. For this reason, the use of these techniques is suggested to increase diagnosis and prognosis. This work seeks to take advantage of available genetic data and ML algorithms in an attempt to gain a better understanding of the risk factors associated with AD.

1.4 Hypothesis, Objectives and Contributions

The main objective of the research is to extend the set of genetic risk factors for AD. The hypothesis behind this aim is that:

> *The use of supervised machine learning methods will identify combinations of SNPs that improve the prediction of Alzheimer's Disease compared to those obtained by univariate methods.*

The hypothesis is supported by ML techniques which have demonstrated their power when managing vast amounts of predictors. Feature selection techniques will be used to extract the most informative attributes and search algorithms coupled with classifiers will build models that maximize the prediction. The set of markers identified will allow to find an extended set of possible genes involved in the susceptibility to suffer from AD.

1.4.1 Specific Objectives

Some of the particular goals that this research includes are:

1. Identify genetic data sources.
2. Get data and put in a proper format.
3. Split data by chromosome to improve computability.
4. Filter features and samples by quality and adequacy.
5. Perform classical univariate analysis.
6. Build machine learning models from top univariate features.
7. Analyze and compare the prediction of models.
8. Analyze and compare the biological content of models.

These objectives would help to answer important questions such as:

1. Is genetic data enough to predict the disease?
2. Does the sample size allow making valid conclusions about the performance of the models?
3. Are the ML methods robust enough to guarantee reproducibility of the results?

4. Do the selected variants within the different models matched?
5. Does the multivariate analysis help to identify new genes?
6. What are the advantages and disadvantages of multivariate methods compared to classical univariate approaches?

1.4.2 Contributions

Below is shown a list of the main contributions of this research. These were established to achieve the objectives and to answer the questions previously raised.

1. It presents an overview of the current state of the art of the use of machine learning algorithms with genetic data to increase predictions of several diseases.
2. Discussions of the advantages of conventional univariate approaches are presented where a pre-filtering stage of variables is included to reduce features size.
3. Two main strategies of the solution are presented: one that is guided by random combinations and another that is based on statistics.
4. For the statistical search strategy, forward selection and backward selection methods are used to generate the combinations that maximize the prediction.
5. For the stochastic search strategy, an approach that couples genetic algorithms with a classifier and a forward selection procedure is used to generate different models and evaluate their performance.
6. LASSO outperformed the other models with an AUC of 0.8005, which is a promising result since Alzheimer's disease has genetic, environmental and lifestyle risk factors and the models in this study were based only on genetic data.
7. A two steps approach modeling poorly classified samples was highly successful.
8. The biological relevance of the markers of each model was analyzed by mapping the genetic region to which they belonged. AD-associated genes were replicated, especially near APOE on chromosome 19.

1.5 Document Structure

The present document has the following structure, Chapter 1 presents an overview of the research as well as the problem faced. Hypothesis, objectives and main contributions are also mentioned. Chapter 2 shows the theoretical bases that support the research and a description of the previous works that have proposed the use of machine learning with genetic data.

Chapter 3 describes the proposed solution and establishes the methodology followed by the investigation to reach its objectives. Subsequently, Chapter 4 presents the database used and the description of the quality control that was carried out. Chapter 5 explains the univariate strategy that was used to pre-filter data and then, Chapter 6 shows the performance of the models obtained with the proposed solution strategies and the different subsets of data used in each one of them. In Chapter 7 the predictors of the different models obtained are analyzed, and the biological interpretation is carried out. Finally, Chapter 8 mentions the conclusions of the research and describes future improvements.

2 | Theoretical Framework

> *"Data starts to drive the operation; it is not the programmers anymore but the data itself that defines what to do next".*
>
> - Ethem Alpaydin, *Machine Learning: the new AI [8].*

This dissertation attempts to identify sets of markers that maximize the prediction of AD through computational methods, so it is essential to examine theoretical principles and other research efforts published. Throughout the chapter, essential concepts about ML will be explained, as well as metrics to evaluate the performance of the algorithms. Additionally, the section describes the previous work done by different authors when incorporating ML techniques in the prediction of several diseases.

2.1 Machine Learning

Machine Learning is a sub-field of artificial intelligence based on statistical and optimization techniques. Its main objective is to build models learning parameters from data to predict outcomes in samples they have not seen before [13]. The learning algorithm is capable of adjusting the parameters of a model by itself to optimize a specific performance user-defined metric.

The term was coined for the first time in 1959 by Arthur Samuel, a pioneer of artificial intelligence research who defined it as a *Field of study that gives computers the ability to learn without being explicitly programmed*. In the last decades, the term has become popular due to its versatility and power [45]. Additionally, the improvements in computing power have eliminated the barriers to handle large amounts of data [61]. Despite the different applications, ML methods allow us to understand the relationship between the samples and their features. Usually, the data is ordered in a table called dataset where the rows correspond to the examples or instances and the columns to the predictors or features. Table 2.1 shows that a GWAS can be represented as a machine learning problem. The markers, SNPs or variants, are the

Machine Learning	GWAS
(Condition) attribute/feature	Genotype
(Condition) attribute/feature	Covariate
Decision attribute	Phenotype
Instance	Individual from the population described by its condition and decision attributes
Training set (in particular training set or test set)	Population sample

Table 2.1: Translation between genome-wide association studies (GWAS) and machine learning terminology. Taken from Ronald M. Nelson et al. (2013) Higher Order Interactions: Detection of Epistasis Using Machine Learning and Evolutionary Computation [45].

features that describe the genotypes of each sample. The phenotype is the decision attribute or target that assigns the label of whether the sample corresponds to a case or a control.

ML has many applications such as image and speech recognition [16], anomaly detection [48], customer segmentation [50], medical diagnosis [58], and gene expression [41]. Figure 2.1 shows an example of an application in genomics, where whole-genome sequences are used to predict locations of transcription start site (TSS) [36].

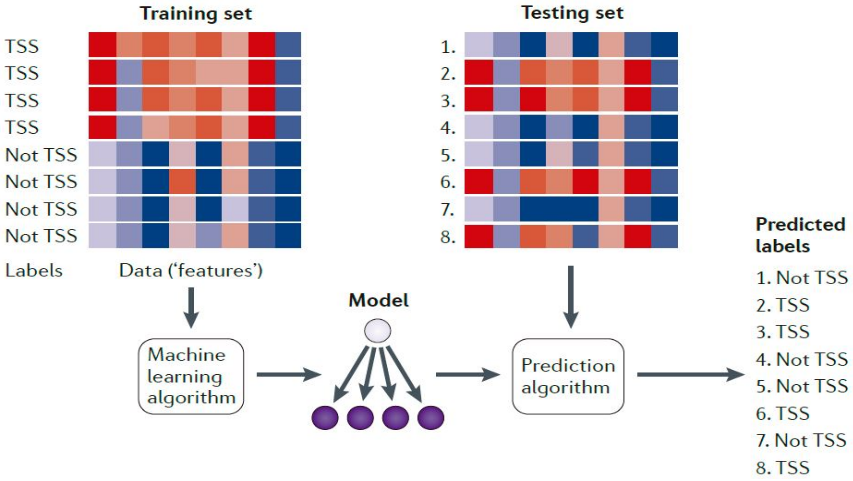


Figure 2.1: Example of a Machine Learning application. Taken from Libbrecht MW and Noble WS (2015) Machine learning applications in genetics and genomics [36].

ML can be divided in two sub-areas depending on the type of problem: supervised learning and unsupervised learning [18]. Supervised learning is when the label of the samples is defined, and the main tasks include regression and classification [36]. On the other hand, unsupervised learning has no target defined and its objective is to reach conclusions with the information that the examples have in common; tasks of this type are known as clustering [8]. In this research, supervised methods will be used for classification.

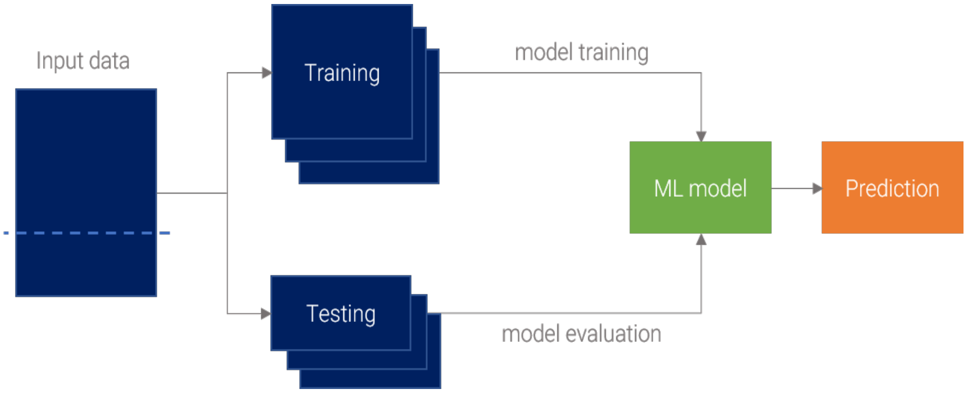


Figure 2.2: Supervised Machine Learning workflow.

2.1.1 Supervised Learning

Supervised learning is used when sample labels are known. Supervised methods build models with labeled instances, and test it in a different dataset that does not have defined labels. Figure 2.2, displays a basic workflow of supervised learning for classification. The process can be split into the following steps:

1. **Input Data:** Raw data goes through a filtering process that eliminates poor quality samples, outliers, missing values, among others. Input data has defined labels and is randomly split into training and testing sets.

2. **Model training:** Machine learning algorithms are used to build models from training data. There is a wide variety of classification methods, including Naive Bayes, Logistic Regression, Random Forest, Support Vector Machines and *k*-Nearest Neighbors.

3. **Model Evaluation:** This phase evaluates the model performance by estimating the error in the testing set and, if necessary, parameters are adjusted to improve model accuracy.

4. **Prediction:** The model built is tested on a new unlabeled dataset and generates a prediction which classifies the unseen instances with the experience achieved in the training phase.

2.1.2 Feature Selection

Feature selection methods face the problem of having a large number of variables by reducing the number of irrelevant or redundant variables and keeping the ones with more relevance with the outcome. By applying feature selection techniques, further computation becomes lighter, providing a better understanding of the problem, and, generally increasing the prediction accuracy [20]. There are different categories according to the criteria used to select features: filter methods, wrapper methods, and embedded methods, which are described in following sections.

Filter Methods

Filter methods are a pre-processing step before the classification approach is selected [22]. Filter techniques are based on variable ranking procedures to order the features by relevance and then use a criterion to filter out the less relevant variants. The relevance may be a statistical test, a correlation function, or any measure of association.

The advantages of variable ranking methods are its simplicity, its computational efficiency and its robustness against overfitting. Overfitting occurs when the classifier model learns the training data too well, having a poor generalization ability. For this reason, this kind of techniques are the most commonly used in GWAS, such as Fisher's exact test and Armitage trend test [47]. However, some of the disadvantages of ranking methods are that they do not consider combinations between variables, a variable may be irrelevant on its own, but joining with others may increase its predictive power. Another disadvantage is that some methods, such as Pearson's correlation criteria, do not discriminate redundant variables, so subsets are not optimal.

Wrapper Methods

Wrapper methods use selection criteria based on classifiers performance when evaluating attribute subsets [20]. To test all the interactions between the markers of a GWAS, 2^n subsets evaluations are required (where n is the number of SNPs and frequently goes from half million to a million) this computation time is unfeasible. For this reason, wrapper methods use exhaustive search strategies, which, despite not finding global optimum, find good local optima in a reasonable time. There exist several search approaches that explore the data while maximizing optimization functions. A general classification by the strategy used is Greedy Search Algorithms and Heuristic Search Algorithms.

Greedy Search Algorithms

Greedy strategies are simple techniques that make optimal choices at every step by evaluating an optimization function in an attempt to find the global solution [20]. Among its advantages stand out computational efficiency and robustness against overfitting. There are two main strategies: Forward Selection and Backward Selection.

- Forward Selection (FS): The algorithm starts with an empty set and chooses the attribute with the highest value for the objective function. After taking the first attribute, the algorithm repeats the process to pick a second feature and attach it to the first feature chosen. This process is repeated until the addition of more features does not increase the classification accuracy.

- Backward Selection (BS): This algorithm follows an approach quite similar to FS but starts with a set that contains all the features and removes one variant at a time until there is no loss in the classifier performance.

Heuristic Search Algorithms

Heuristic search algorithms evaluate different subsets generated by search methods with the goal of finding the one that optimizes the objective function [20]. Evolutionary Computation (EC) techniques have been used to face the problem of searching optimal solutions in high dimensional spaces since their objective is to find the best subset of features in the least searching time. One of the most used EC algorithms is Genetic Algorithms [4].

Genetic Algorithms (GA) are algorithms inspired in biological processes of evolution. Figure 2.3 displays the main steps in a GA. The *initialization* step generates a population of random solutions, where each solution contains a subset of predictors known as chromosomes. The following step does an *evaluation* of the solutions using a fitness function, which measures the performance of each solution and uses it as a measure of reproduction possibilities. In the *selection* phase, the solutions with a higher fitness evaluation are chosen for reproduction [14]. *Reproduction* step allows selected solutions to be cloned or recombined. The *mutation* process is applied to some solutions creating unique variations. Finally, new solutions replace the old population of solutions, and the process is repeated for all generations. The algorithm stops when a defined fitness is reached or when the number of iterations ends.

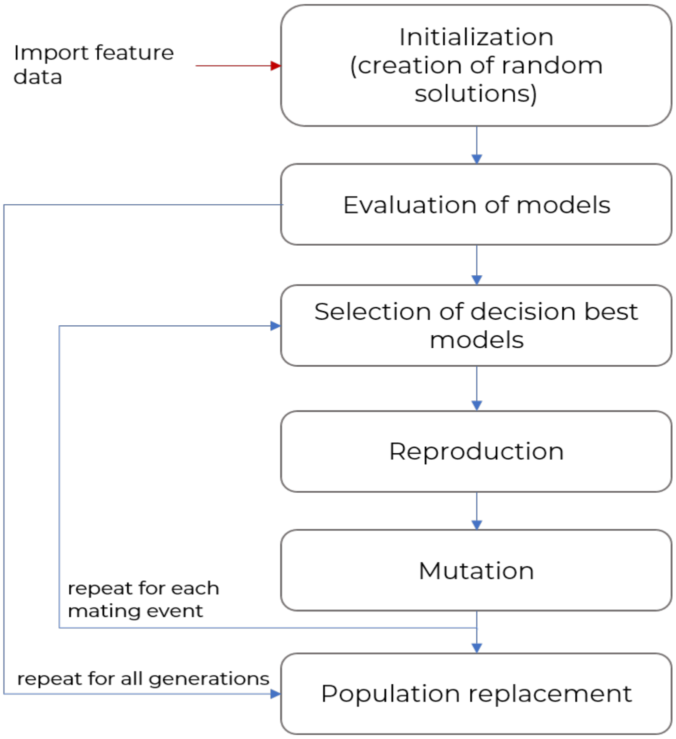

Import feature data
Initialization (creation of random solutions)
Evaluation of models
Selection of decision best models
Reproduction
Mutation
repeat for each mating event
repeat for all generations
Population replacement

Figure 2.3: Basic steps of a Genetic Algorithm. Adapted from Ronald M. Nelson et al. (2013) Higher Order Interactions: Detection of Epistasis Using Machine Learning and Evolutionary Computation [45].

In conclusion, wrapper methods are a good alternative as a search strategy in problems with high dimensionality. However, one of its weaknesses is the computation time since they perform many calculations to obtain each subset of features. Also, the excessive use of accuracy estimation may cause overfitting, since a subset of features with high accuracy but poor predictive ability could be selected [26].

Embedded Methods

Embedded methods are characterized by including a feature selection process in the training stage of the model [22]. These methods take up an approach similar to the wrapper methods since they also look for solutions in the whole search space guided by the performance measurement of a classifier, but reducing the computation time. A well-known method is LASSO, and it has been applied recently in many genetic association studies [47].
Least Absolute Shrinkage and Selection Operator (LASSO) is a regression-based method originally designed to reduce overfitting that minimizes the absolute sum of the coefficients using L1-regularization [52]. The penalty causes many of the weakly associated coefficients to be zeros, consequently, LASSO performs a kind of feature selection [18]. The LASSO coefficients $\hat{\beta}^{L}_{\lambda}$, minimize the following function:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \tag{2.1}$$

where $\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2$ is the Residual Sum of Squares and $\lambda\sum_{j=1}^{p}|\beta_j|$ is the norm of the coefficient vector β multiplied by the non-negative tunning parameter λ.
LASSO improves prediction accuracy and generates translating models [43].

2.1.3 Classifiers

A classifying task consists of generating a model that assigns a class to an unseen sample according to their features. There are many classification methods, each with different characteristics and suitable for diverse problems. This section provides some of the best-known classification approaches.

Support Vector Machines

Support Vector Machines (SVM) is a classification method that maximizes the margin between samples from two classes by an efficient separation of hyperplanes in high dimensional feature space [57]. The decision boundary is a hyperplane of the form:

$$\mathbf{w}\mathbf{x} + b = 0 \tag{2.2}$$

where $\mathbf{w}$ is a weight vector of linear combinations of training patterns, $\mathbf{x}$ is the input vector and b is the bias. The support vectors are the points that rely on the margin. The distance between planes $\mathbf{w}\mathbf{x} + b = +1$ and $\mathbf{w}\mathbf{x} + b = -1$ is $\frac{2}{||\mathbf{w}||}$. To maximize the margin, $||\mathbf{w}||$ is minimized. Figure 2.4 illustrates optimal separating hyperplane of SVM.

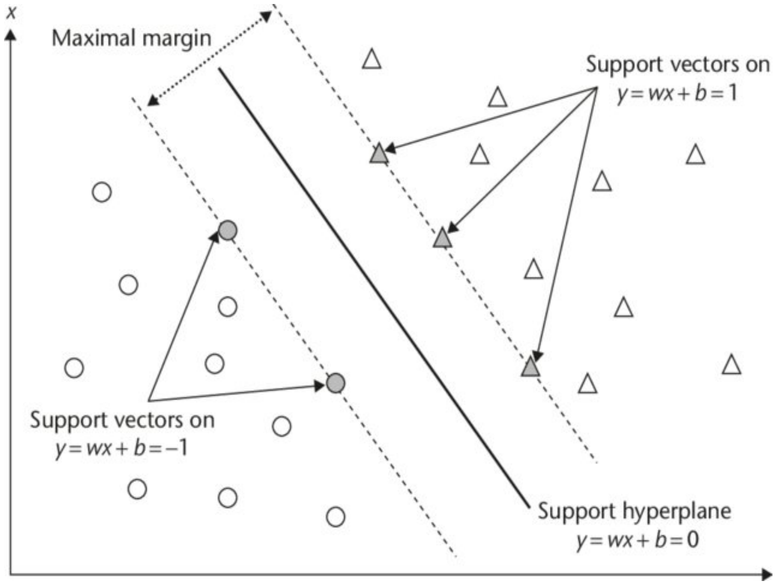


Figure 2.4: Optimal separating hyperplane of support vector machine (SVM). Taken from Huang, Y. (2009) Computer-aided Diagnosis Using Neural Networks and Support Vector Machines for Breast Ultrasonography [57].

SVM uses kernels to map the input data higher dimensional feature space. The most common kernels are polynomial, radial, and ANOVA. The advantages of SVM include a unique solution, flexibility due to the use of kernels and robustness when the parameters are correctly chosen [1]. However, the parameter estimation is considered as a disadvantage because some values can lead to good classification accuracy in one problem, but may drive to a poor performance in another.

2.1.4 Validation Methods

Once a model is built, it is essential to validate its generalization performance in data different from the training set. One of the most used validation techniques in machine learning is cross-validation.

Cross-Validation

Cross-validation is a method that divides the initial dataset into a set for training and a set for testing. Its objective is to estimate efficiently the ability of a model to predict in different samples to those that were used in the construction of the model, preventing overfitting and bias. Figure 2.5 illustrates the K-fold cross-validation method.
K-fold cross-validation splits the original dataset into k subsets called folds. In each iteration, one fold is chosen for testing and, the remaining $(k-1)$ folds form the training set that is used to build the model. Once the performance of all the k-folds is obtained, the results are

averaged to have a total accuracy of the model.

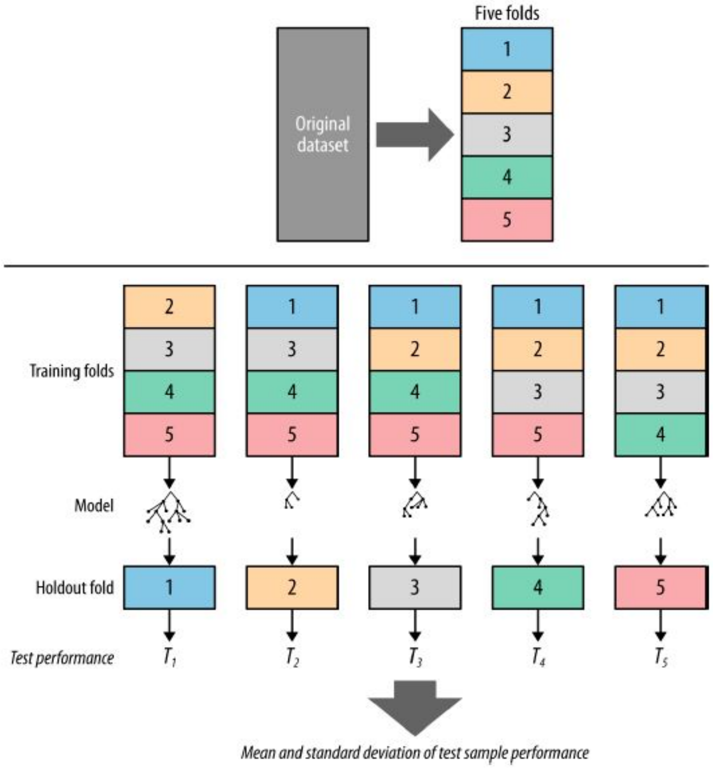


Figure 2.5: K-fold cross-validation method. Taken from Provost, F. and Fawcett, T. (2013) Data Science for Business [18].

2.2 Performance Metrics

Once the models are built, metrics are required to evaluate their prediction performance. In binary classification tasks, a confusion matrix is a useful tool that allows visualizing the performance of a classifier.

Table 2.2 shows a confusion matrix, where the rows represent the predicted class while the columns represent the current class. The terms True Positive (TP) and True Negative (TN) are the cases that were correctly classified. On the other hand, the False Positives (FP) are those examples that were false and were predicted as positive and the False Negatives (FN) are those that were predicted as negative but were positive [32].

Many performance measures are derived from the confusion matrix. The most used metrics in practice are accuracy, error rate, sensitivity, and specificity. These metrics are calculated with the following formulas:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.3}$$

	Actual Positive Class	**Actual Negative Class**
Predicted Positive Class	TP	FP
Predicted Negative Class	FN	TN

Table 2.2: Structure of a confussion matrix. TP: True Positives, FP: False Positives, FN: False Negatives and TN: True Negatives.

Accuracy computes the percentage of correct predictions over total instances. It is a widely used metric because it is computationally inexpensive and allows to see overall effectiveness of a classifier. However, accuracy should not be used when classes are unbalanced [21].

$$Error \quad rate = \frac{FP + FN}{TP + FP + TN + FN} \tag{2.4}$$

The error rate, unlike accuracy, measures the ratio of incorrect predictions over the total number of evaluated instances.

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.5}$$

Sensitivity measures the proportion of positive samples that were correctly predicted.

$$Specificity = \frac{TN}{TN + FP} \tag{2.6}$$

Specificity measures the proportion of negative examples that were correctly classified.
The performance of a classification model can also be evaluated through an analysis of Receiver Operating Characteristic (ROC) curves. In some real applications, such as in medicine, case-control discrimination is rarely perfect [33], as shown in figure 2.6.

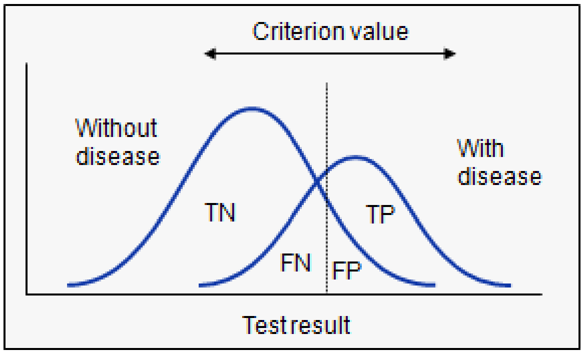


Figure 2.6: Distribution of test results in a medical example. Taken from MedCalc Software (2019) ROC curve analysis [37].

Each criterion value will include some well-classified cases (TP), but also some FN. Conversely, it will cover some instances correctly labeled as controls (TN), but some misclassified samples (FP). The ROC curves allow visualizing said thresholds plotting the true positive rate against the false positive rate. Each classifier is represented by a point (FP, TP) in the ROC curve, as shown in figure 2.7. A model will have better performance if the curve is closer to the upper left side of the graph.

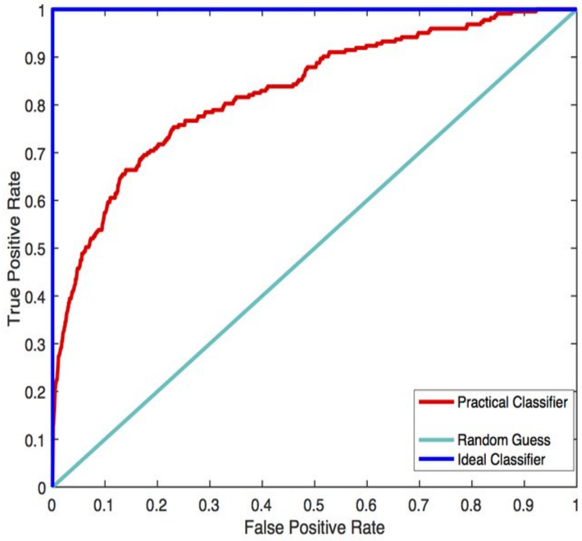


Figure 2.7: The Receiver Operating Characteristic (ROC) curves. Taken from Chen, Hongge (2017) Novel machine learning approaches for modeling variations in semiconductor manufacturing [11].

The area under the ROC curve (AUC) is a statistical criterion used to measure the ranking quality of a classifier and is calculated as follows:

$$AUC = \frac{1 + TPrate - FPrate}{2} \tag{2.7}$$

Where TP_{rate} is the $sensitivity$ and FP_{rate} is $1 - specificity$. The higher the AUC, the better the model is considered at discriminating between classes.

2.3 Previous Works

As mentioned above, there is a need to understand the genetic factors associated with various complex diseases to have a better understanding and prevent them as well as generate specialized treatments. Despite the success of the univariate analysis by identifying some genes as risk factors for several diseases, the problem of missing heritability remains. One way to deal with this problem is through multivariable machine learning techniques that consider the interaction between markers. The application of ML techniques based on genetic data has been popularized to identify risk factors associated with several diseases, since they have proven

Author	Year	Title	Methods
Zhi Wei et al.	2009	From Disease Association to Risk Assessment: An Optimistic View from Genome-Wide Association Studies on Type 1 Diabetes	FS: Univariate method; MB: SVM radial AUC=0.84, LR AUC=0.80

Zhi Wei et al.	2013	Large Sample Size, Wide Variant Spectrum,and Advanced Machine-Learning
		Technique Boost Risk Prediction for Inflammatory Bowel Disease
			FS: Univariate method;
			MB: CD LR-L1 AUC= 0.86;
			CD SVM AUC=0.862; CD GBT AUC= 0.802

Lee et al.	2018	Machine Learning on a Genome-Wide Association Study to Predict Late Genitourinary Toxicity Following Prostate Radiotherapy	FS: Chi-squared test, LR; MB: PRFR AUC=0.70
Maciukiewicz et al.	2018	GWAS-based machine learning approach to predict duloxetine response in Major Depressive Disorder	FS: LR, LASSO; MB: Linear SVM ACC=0.66, CRT ACC=0.57

Table 2.3: Summary of ML approaches based on GWAS data for several diseases and traits over the last decade. CD: Crohn's Disease, FS: Feature Selection, MB: Model Building, SVM: Support Vector Machines, LR: Logistic Regression, PRFR: Pre-conditioned Random Forest Regression, ACC: Accuracy, AUC: Area Under the Curve.

to be robust when the problem involves hundreds of thousands of predictors. The following paragraphs present some studies that have applied several ML methods based on GWAS data to predict different diseases during the last decade. A summary is given in table 2.3.

From Disease Association to Risk Assessment: An Optimistic View from Genome-Wide Association Studies on Type 1 Diabetes - 2009

In this article, Wei et al. [58] demonstrate that SVM and logistic regression (LR) obtain higher predictions when entering large amounts of polymorphisms as opposed to taking only a few previously associated markers. They used the Type 1 Diabetes (T1D) GWAS dataset from the WTCCC containing 1,963 cases and 1,480 controls. They generated several models with different thresholds that included p-values from 1×10^{-3} to 1×10^{-8}. The validation of the model was done with 5-fold CV and 80% of the dataset for training and 20% for testing. Their results yielded that SVM outperforms LR with an AUC ~ 0.90 with a p-value cutoff of 1×10^{-5} which corresponded to 399-443 SNPs.

They also validated the model prediction generated with WTCCC-T1D in two independent datasets: CHOP / Montreal-T1D (1,008 cases and 1,000 controls) and GoKind-T1D (1,529 cases and 1,458 controls). The results again showed better performance by SVM with an AUC of 0.83 in the CHOP/Montreal-T1D dataset with a p-value threshold of 1×10^{-5} (478 SNPs) while under the same conditions LR achieved an AUC of 0.622. Evaluating prediction in Gokind-T1D dataset with a p-value cutoff of 1×10^{-5} (409 SNPs) showed similar results with AUC scores of 0.84 and 0.805 for SVM and LR respectively. Finally, they built a model with only 45 SNPs previously associated with T1D; when validating with CHOP / Montreal-T1D, SVM obtained an AUC of 0.65 and LR an AUC of 0.68 while with GoKind-T1D with SVM achieved an AUC of 0.66 and LR an AUC of 0.68. The authors conclude that the use of ML methods has the potential to explore large amounts of data from GWAS reaching high prediction levels that help to have a better understanding of the prevention and treatment of diseases.

Large Sample Size, Wide Variant Spectrum, and Advanced Machine-Learning Technique Boost Risk Prediction for Inflammatory Bowel Disease - 2013

In this study, Wei et al. [59] propose a feature selection strategy in two steps: univariate filtering followed by a penalized regression method for risk assessment in Crohn's Disease (CD) and Ulcerative Colitis (UC). The genotype data used came from the International IBD Genetics Consortium's Immunochip project that contains 17,379 CD cases, 13,458 UC cases, and 22,442 controls and they formed three sub-datasets for pre-selection, training, and testing for each disease. For CD, they performed univariate association tests with the pre-selection subset by filtering SNPs with p-values$> 1 \times 10^{-4}$ and minor allele frequency (MAF) < 0.01, obtaining 10,799 SNPs. A logistic regression method was used with L1-regularization with the filtered markers, adjusting the lambda value through 10-fold CV and the model achieved a training AUC of 0.864 with a predictive model of 573 SNPs. For UC they followed the same strategy, a filtered set of 6,968 SNPs was used as input to the classifier, achieving an AUC of 0.83 and a predictive model of 366 SNPs. When testing the models with the third subset that was initially made, the results showed an AUC of 0.864 for CD, while an AUC of 0.826 for UC.

The authors analyzed the possible causes of the model performance and carried out experiments modifying the sample size, concluding that the larger the sample, the better the prediction. They also compared the linear strategy against other non-linear methods, SVM with radial kernel and Gradient Boosted Trees (GBT). SVM obtained an AUC of 0.862 for CD and an AUC of 0.826 for UC, while GBT had lower performance with an AUC of 0.802 and 0.782 for CD and UC, respectively. The writers conclude that the use of a double feature selection strategy makes the model selection more feasible since the dimensionality is drastically reduced and high predictive power is achieved as long as a sufficiently large sample is available.

GWAS-based machine learning approach to predict duloxetine response in major depressive disorder - 2018

In this article, Maciukiewicz et al. [30] present a machine learning strategy with a feature selection stage followed by a model optimization stage based on GWAS data for the prediction of duloxetine in patients with major depressive disorder (MDD). The dataset used consisted of 450 patients who had been diagnosed with MDD and 571,054 SNPs genotyped at The Center for Applied Genomics (TCAG). Quality control removes samples with excessive heterozygosity, excessive relatedness, non-caucasian ancestry and patients treated with placebo; reducing the number of samples to 186 patients. On the other hand, quality control of the genotypes included filtering by minor allele frequency$> 1\%$, Hardy-Weinberg equilibrium ($p > 1 \times 10^{-7}$), genotype call rate$> 98\%$ and individual missingness <10% resulting in 285,950 SNPs. They also performed a whole-genome imputation using IMPUTE v2.2, which resulted in 1,924,595 SNPs. Figure 2.8 shows the methodology used to build a model that predicts the response and remission to the antidepressant. They implemented a nested cross-validation strategy, in the outer loop they used 5-fold CV and performed feature selection using two techniques: logistic regression to filter clinical variables with little impact on the response/remission and lasso to extract the most promising predictors. For the construction of the models, they used two classifiers: linear SVM and Classification Regression Trees (CRT). An inner loop of 10-fold

CV was used for parameter tuning. Two types of models were built: 1) based on the originally genotyped gene variants only and 2) based on the originally genotyped and imputed genetic variants.

The results showed a low performance of both algorithms to predict the response to duloxetine. For the genotypes with imputed data, SVM achieved an accuracy of 0.64 and CRT an accuracy of 0.57 whereas for only the imputed data the accuracy was 0.66 for SVM and 0.55 for CRT. The maximum value of sensitivity (0.89) was obtained by SVM with genotyped data only. However, it had a very poor specificity of 0.09. On the other hand, for the prediction of remission, there was also little predictive power with a maximum accuracy of 0.52, a sensitivity of 0.58 and specificity of 0.46 reached by SVM with genotyped data with imputed data. The authors conclude that the poor performance to predict the response may be due to the unbalanced classes. They also add that it was a purely computational study but that adding biological information could increase the accuracy of the models.

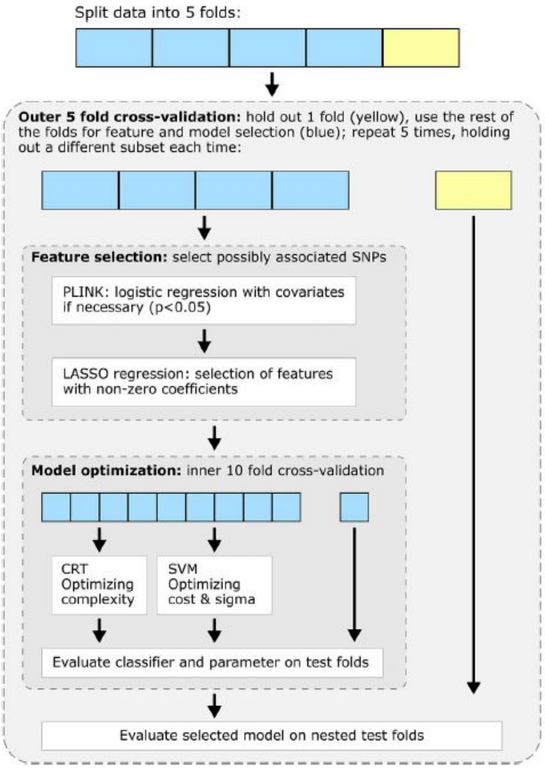

Split data into 5 folds:
Outer 5 fold cross-validation: hold out 1 fold (yellow), use the rest of the folds for feature and model selection (blue); repeat 5 times, holding out a different subset each time:
Feature selection: select possibly associated SNPs
PLINK: logistic regression with covariates if necessary (p<0.05)
LASSO regression: selection of features with non-zero coefficients
Model optimization: inner 10 fold cross-validation
CRT Optimizing complexity
SVM Optimizing cost & sigma
Evaluate classifier and parameter on test folds
Evaluate selected model on nested test folds

Figure 2.8: Methodology of the prediction model proposed for the response/remission to an antidepressant. Taken from Maciukiewicz et al. (2018) GWAS-based machine learning approach to predict duloxetine response in major depressive disorder [30].

Machine Learning on a Genome-Wide Association Study to Predict Late Genitourinary Toxicity Following Prostate Radiotherapy - 2018

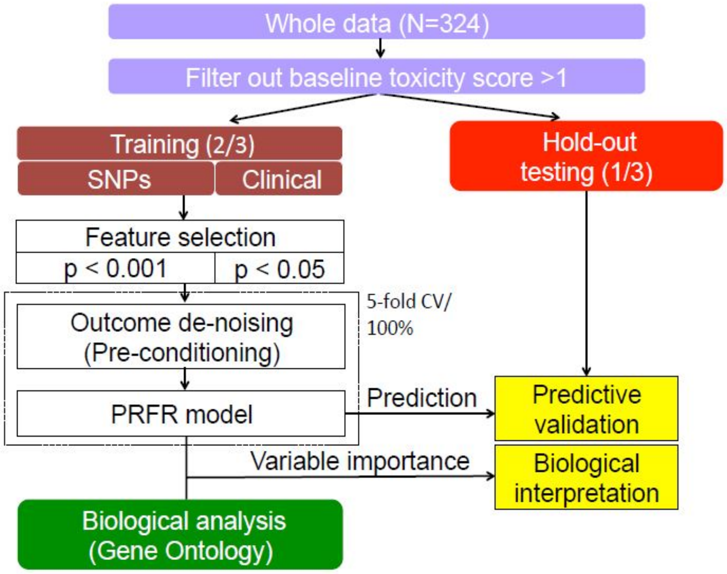


Figure 2.9: Flowchart of the methodology proposed to predict the incidence of GU toxicity symptoms following Radiotherapy. Taken from Sangkyu et al. (2018) Machine Learning on a Genome-Wide Association Study to Predict Late Genitourinary Toxicity Following Prostate Radiotherapy [46].

Sangkyu et al. [46] propose the use of Pre-conditioned Random Forest Regression (PRFR) to identify patients with high risk of late genitourinary (GU) toxicity using, GWAS data. The cohort consists of 324 prostate cancer patients who presented urinary symptoms (incomplete emptying, frequency, intermittency, urgency, weak stream, straining, and nocturia) two years after radiotherapy and 606,563 SNPs before quality control. The dataset was divided into $2/3$ for training and $1/3$ for testing. A pre-filtering stage was performed using a chi-squared test (p-value <0.001) for the SNPs that passed the quality control. Continuous clinical variables were evaluated using logistic regression (p-value <0.05). The pre-conditioning step transforms the original categorical outcomes into continuous outcomes through logistic regression coupled with principal components, maximizing the correlation between them. The model was constructed using Random Forest (RF) with the pre-conditioned outcomes in a 5-fold CV scheme repeated 100 times with randomized fold configuration. The performance of the PRFR model was compared against (1) RF without pre-conditioning, (2) LASSO, (3) pre-conditioned LASSO, and the PRFR model re-trained with a number of SNPs with (4) the top 50%, (5) the top 75% of SNPs based on Variable Importance Measures (VIM), and (6) ethnicity-based model. The predictions of the PRFR model varied among the symptoms with the highest classification performance predicting weak stream symptom. For the 5-fold CV scheme, the AUC was 0.67 while building the model using all the training data the AUC was 0.70. The authors conclude that the PRFR approach significantly outperformed the conventional RF and the linear models (p <0.001). The use of bioinformatics tools and ranking the importance of the SNPs in the PRFR model allowed to identify biological processes and proteins involved in radiation injury.

2.4 Summary

This chapter introduced the basic concepts used in the present study. The most used feature selection strategies were described including filter, wrapper, and embedded methods. Furthermore, the chapter presents one of the most used classifiers in the area of medicine as well as validation techniques and metrics that are used to compare the performance of different methods. Finally, some relevant studies are presented highlighting the machine learning methods that have been used in the last decade to predict diseases based on GWAS data.

3 | Solution Model and Methodology

3.1 Solution Model

The state-of-the-art review of ML methods based on GWAS data as a prognosis tool showed an overview of the strategies that have been used in recent years to build predictive models that help identify genes associated with various diseases or treatments. The characteristics that stand out from previous works are:

1. All the previously mentioned approaches are based on GWAS data. Samples and markers go through a quality control procedure to decrease the number of false positives and false negatives.

2. Each article implements a dimensionality reduction strategy, either with a pre-filtered stage by an univariate method or by feature selection strategies with penalized regression.

3. For the model building, different approaches are tested, including logistic regressions with penalization, Support Vector Machines with different kernels, and decision trees. The model performance varies according to the disease or treatment, number of samples and number of markers.

This research takes as a starting point the previously used methodologies and proposes a solution that consists of two phases.
First, a pre-selection of features is done to reduce the dimensionality. Second, the model building takes two approaches: wrapper and embedded methods. The wrapper methods include greedy search algorithms (through combinations dependent on rankings and strategies of Forward Selection and Backward Selection for the construction of the model) and heuristic search algorithms (through random combinations generated by genetic algorithms coupled to a classifier). The embedded method used is LASSO since it includes a feature selection previous to the model building. In addition to the analysis of the model, the results include a biological interpretation.

3.2 Methodology

The methodology carried out during this research is illustrated in figure 3.1. The programming was done in the statistical language R [42]. The model building included the use of R packages generated by members of the Bioinformatics research group of Tec de Monterrey.

1. **Genotyped Data**
   The first step was to identify a genotypic database of individuals with Alzheimer's disease. Subsequently, the data is pre-processed including the coding of the genotypes and the creation of matrices by chromosome *(SNPs x samples)*.

2. **Quality Control**
   Genotype matrices will pass through quality control procedures to filter out markers and samples that could increase false positives and false negatives in the association of AD.

3. **Dimensionality Reduction by Feature Pre-Selection**
   The use of univariate techniques is proposed as a pre-filter to reduce the number of predictors. The use of a univariate filtering method had two main objectives, the first one is the efficient computation time analyzing half a million of features and the second one is the replication of the study presented in [17] as a positive control and to verify the correct use of the data. This stage will generate two datasets that will be used by the ML methods: (1) top 100 high-quality SNPs and cases/controls unrelated and, (2) high-quality SNPs with a p-value less than 1×10^-3 and cases/controls unrelated.

4. **Model Building**
   For the model building, two main approaches are proposed: wrapper and embedded methods.

   (a) Wrapper Methods

      i. Greedy Search Algorithm
      The use of FRESA.CAD [5], an R package is proposed since one of its objectives is to assist in the discovery of new features in the health-related area. The supervised method Bootstrapping Stage-Wise Model Selection (BSWiMS) will be used since it builds a statistical model that maximizes the prediction of AD. The algorithm consists of five phases: Univariate Filter, Bootstrapped Forward Selection, Frequency-based Forward Selection, Bootstrapped Backwards, Elimination and Model Bagging.

      ii. Heuristic Search Algorithm
      For the heuristic method, GALGO [53] package will be used since it incorporates genetic algorithms as a search strategy for sets of features coupled to a classifier. GALGO has previously been used in biomarker discovery using gene expression profiling data showing a good performance. In this study, GALGO will be evaluated in a different context with polymorphism data.

(b) Embedded Method
The literature reviewed shows LASSO as a technique with the ability to generate a predictive model since it incorporates a feature selection stage before model training. For this reason, it was decided to use this method to build a prediction model of AD.

i. Misclassified Re-Modeling and Mixture Models
A profound analysis of the samples that are classified incorrectly in the LASSO models is proposed since this strategy seeks to increase the performance of the models by including information from a subgroup of individuals.

The performance of the models will be assessed using accuracy, sensitivity, specificity, and AUC metrics.

5. **Model Analysis and Biological Interpretation**
The last stage is the biological analysis of the obtained models. The markers of each model will be mapped to their corresponding genes to verify their biological relevance.

3.3 Summary

This chapter explained the solution strategy that is taken in this study to fulfill the objective of generating a predictive model for Alzheimer's disease inspired by the main features of the methods used in the literature by applying ML techniques with GWAS data. The methodology of this research was described in five main stages: (1) obtain genotypic data, (2) quality control procedures, (3) dimensionality reduction, (4) model building, and (5) model analysis and biological interpretation.

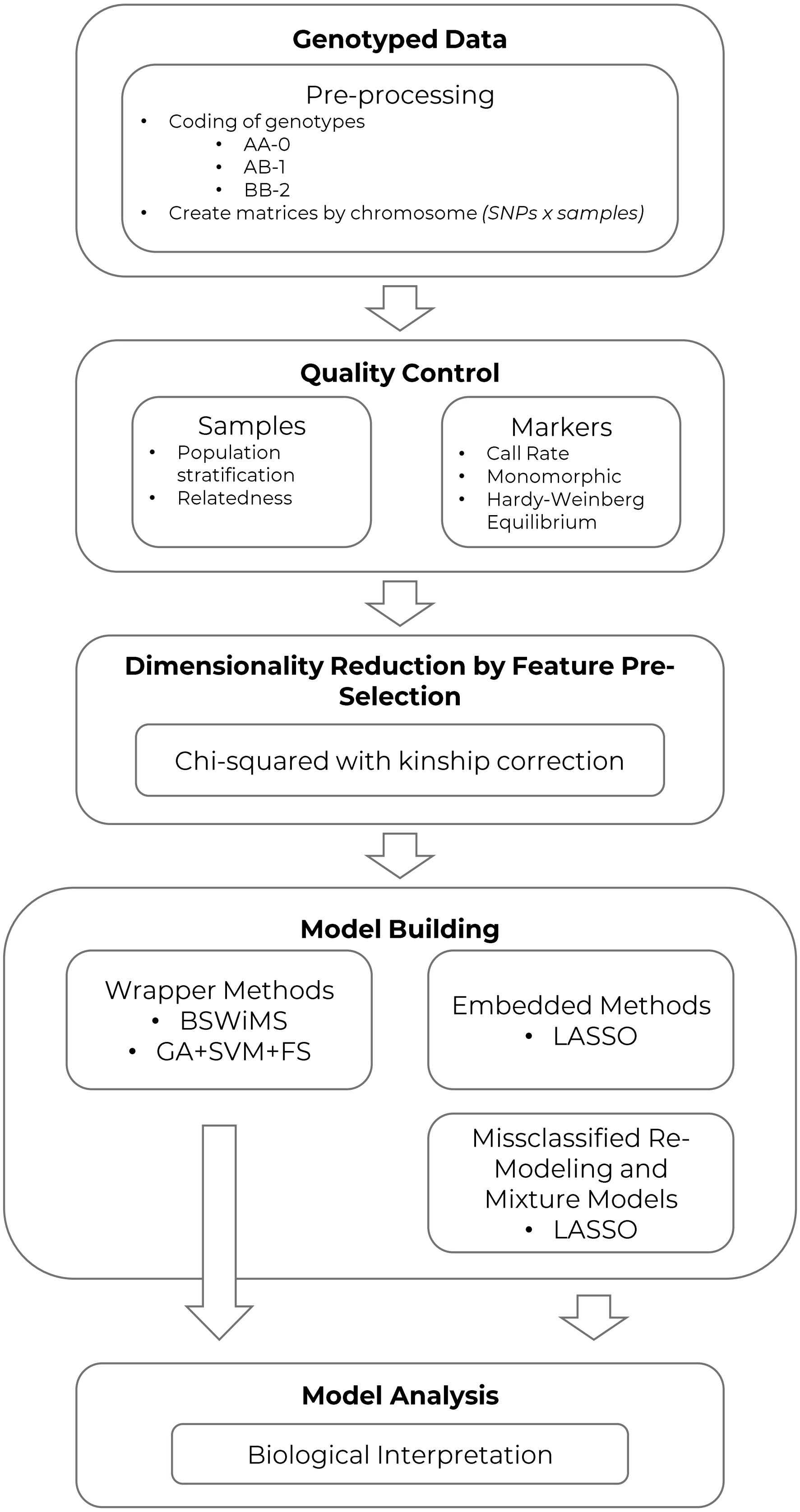

Genotyped Data
Pre-processing
• Coding of genotypes
• AA-0
• AB-1
• BB-2
• Create matrices by chromosome *(SNPs x samples)*
Quality Control
Samples
• Population stratification
• Relatedness
Markers
• Call Rate
• Monomorphic
• Hardy-Weinberg Equilibrium
Dimensionality Reduction by Feature Pre-Selection
Chi-squared with kinship correction
Model Building
Wrapper Methods
• BSWiMS
• GA+SVM+FS
Embedded Methods
• LASSO
Missclassified Re-Modeling and Mixture Models
• LASSO
Model Analysis
Biological Interpretation

Figure 3.1: Methodology proposed to generate a prediction model of Alzheimer's disease with ML methods based on GWAS data.

4 | Data and Pre-processing

The data used in this project correspond to the *National Institute on Aging - Late-Onset Alzheimer's Disease Family Study: Genome-Wide Association Study for Susceptibility Loci (phs000168.v2.p2)* and were requested from the National Center for Biotechnology Information (NCBI) through the Genotypes and Phenotypes database (dbGaP) [15], who approved access in approximately one and a half months. The authorized data contained $\sim$ 200 GB of compressed files, including comma-separated values (CSV) files with genotypic information per individual, CSV files with phenotype information (sex, age, race, case/control, among others) of all individuals and a description of the study.

The study has four groups totaling 5,220 individuals and 620,901 SNPs. Table 4.1 shows in more detail the number of subjects per group. The cases were diagnosed with Definite AD by neuropathological criteria or Probable AD by NINCDS-ADRDA criteria and the age at diagnosis should be equal to or greater than 60 years. The controls were defined as individuals without manifestations of cognitive deterioration or memory loss through neuropsychological tests in addition to not having a history of psychiatric or neurological disorders. More details about the study can be found in [15].

Study Name	Distribution set	Cases	Controls	Other	Total Individuals
General Research Use (GRU)	phs000168.v2.p2.c1	1266	1279	462	3007
Non Profit Use (NPU)	phs000168.v2.p2.c2	12	13	3	28
Disease-Specific Alzheimer Disease (DS-ALZ)	phs000168.v2.p2.c3	719	780	116	1615
Disease-Specific Alzheimer Disease, NPU (DS-ALZ-NPU)	phs000168.v2.p2.c4	323	172	75	570

Table 4.1: Information of individuals by consent group. Taken from dbGaP (2018) National Institute on Aging - Late-Onset Alzheimer's Disease Family Study: Genome-Wide Association Study for Susceptibility Loci [15].

4.1 Pre-processing

In the data pre-processing step, the four groups were combined to form only two groups: cases and controls. Data matrices were created for each of the 22 autosomal chromosomes (where the rows corresponded to the SNPs and the columns to the samples) to lighten the handling of the data in our computers. The number of markers in each matrix varies by chromosome, being 600,470 the total of SNPs. The complete matrices have a dimension (600,470 x 2,319) for cases and (600,470 x 2,242) for controls. The genotypes were coded as 0 when there was a dominant homozygote, 1 for a heterozygote and 2 for a homozygous recessive. Table 4.2 illustrates an example of a genotypes matrix.

	1	**2**	**3**	...	**4,561**
rs12045644	2	1	2	0	2
rs12045674	0	0	0	0	0
rs12045689	1	0	0	0	0
rs12045736	1	1	1	0	2
rs12045750	2	1	1	2	1
rs12045759	2	2	2	1	1
rs12045777	0	0	0	1	1
rs12045781	1	1	2	0	2
rs12045786	0	0	0	0	0

Table 4.2: Genotypes Matrix with markers per rows and individuals per columns *(600,470 SNPs x 4,561 samples).*

4.2 Quality Control

As previously mentioned, Genome-Wide Association Studies (GWAS) measures and analyzes DNA sequence from the human genome to identify genetic risk factors for diseases that are common in the population. The success of a GWAS to detect true genetic associations depends on the quality of the data [49]. For this reason, the data must pass through a Quality Control (QC) to decrease the false positive and false negative associations. The most common QC procedures include identifying poorly genotyped individuals and markers, excessive relatedness, removing duplicate samples, discordant sex information, ancestry correction, among others. The following sections will detail the process.

4.2.1 Sample Quality

The initial sample consisted of 5,220 individuals. The first filtering step was to remove those participants with an intermediate or unknown phenotype (n = 659) that has no assignment of disease or healthy for AD and duplicated samples, which resulted in 4,561 individuals: 2319 cases and 2242 controls. Figure 4.1 summarizes the quality control (QC) procedures for sample assessment.

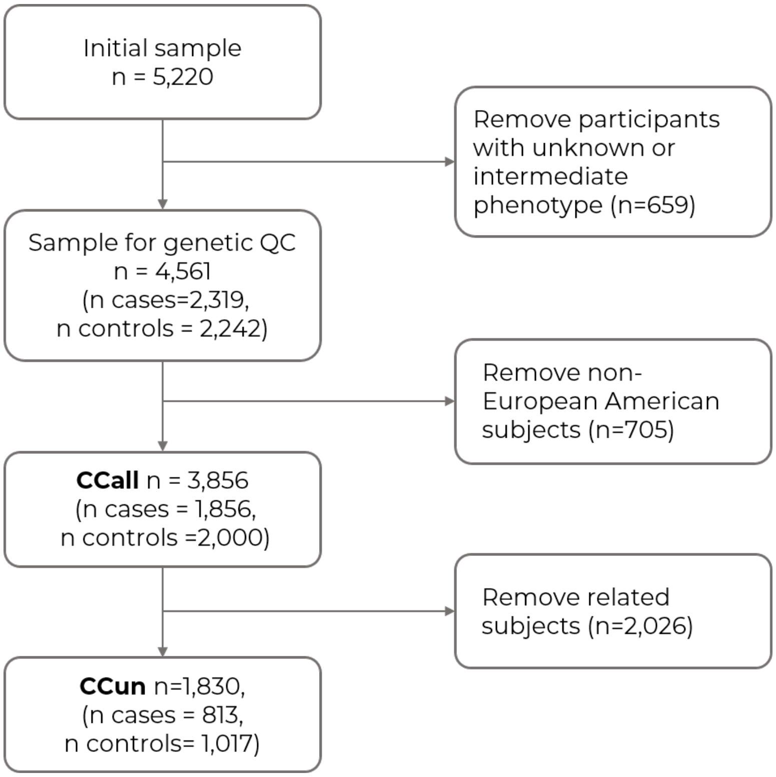


Figure 4.1: Quality Control (QC) procedures for sample assessment.

Population Stratification

The population stratification refers to differences in allele frequencies between cases and controls due to ancestry diversity, which can cause spurious results in association studies [49]. One technique to avoid population stratification is to take individuals from the same ethnic group. In this way, the associations are related to the disease and not to the ancestry. This research faces the problem of population stratification with the EIGENSTRAT method [7] shown in figure 4.2. This approach applies Principal Components Analysis to genotype data to estimate the continuous axes of genetic variation and calculates ancestry-adjusted genotypes and phenotypes by computing residuals of linear regression. Finally, a statistical test is performed using the ancestry-adjusted genotypes and phenotypes.

The filtering by population stratification was performed separately in 2,319 cases and 2,242 controls. Moreover, 100,000 high-quality SNPs were randomly taken as suggested in previous studies [49]. The randomly chosen markers were taken proportionally to the chromosome size, as shown in table 4.3.

The data was processed with the AssocTests package [54] of the statistical software R [42]. Image 4.3 shows the first two principal components of the sample, where the red dots correspond to the European-Americans group. This analysis filtered out 705 individuals of non-European American ancestry. The Euro-American population sample denoted from now on by **CCall** remained in 3,856 individuals of which 1,856 were cases and 2,000 controls.

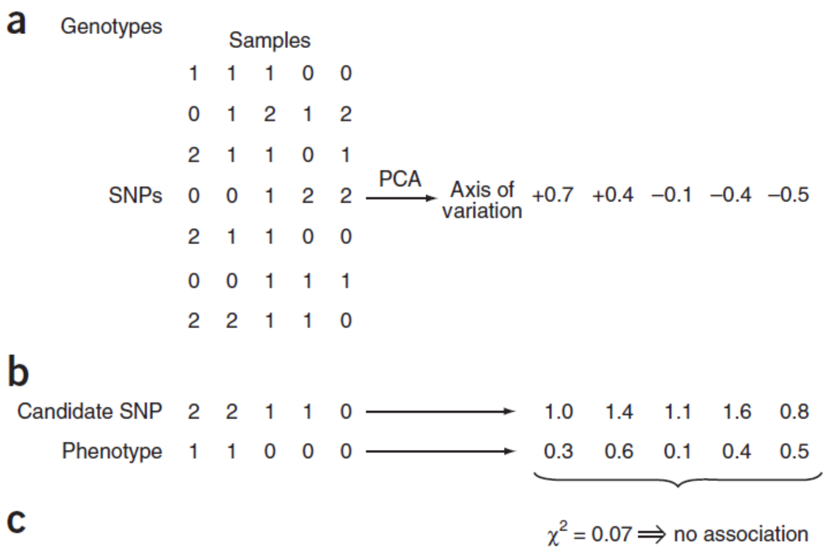
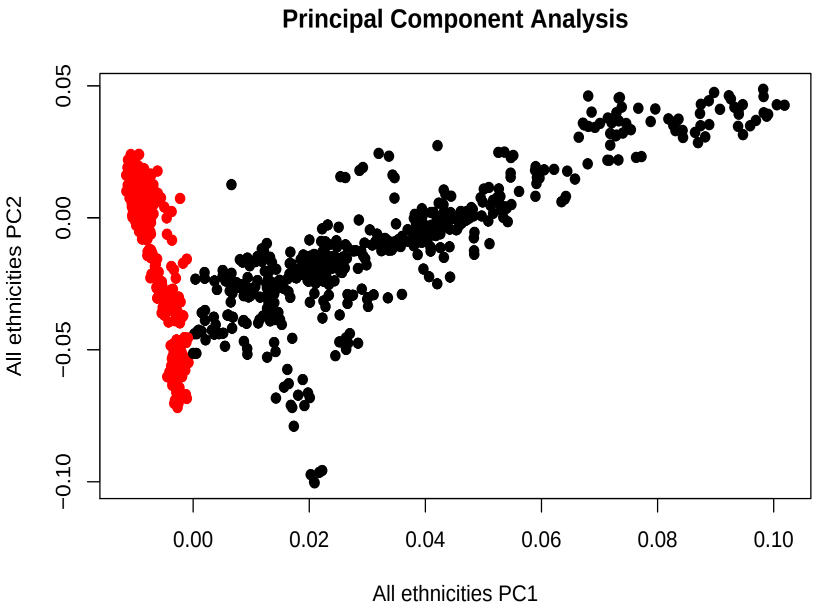


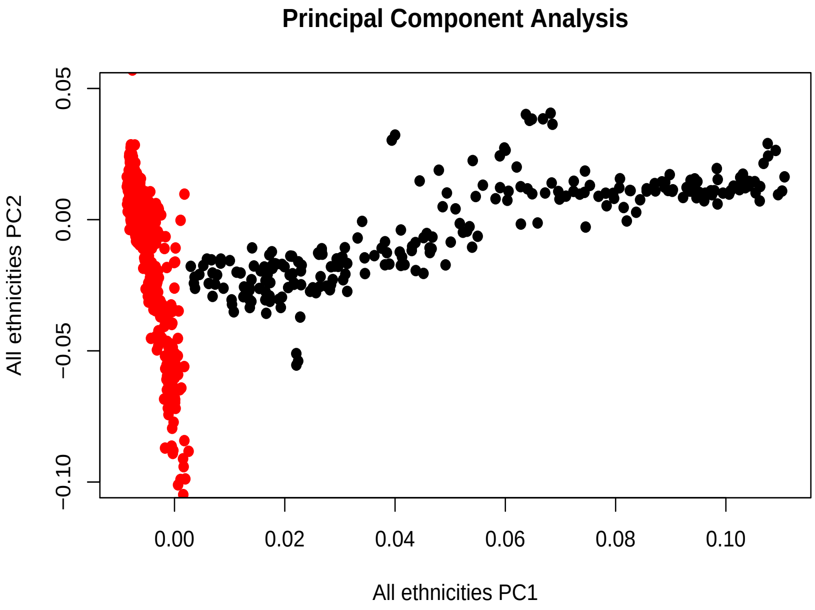
Figure 4.2: EIGENSTRAT algorithm. Taken from Price et al. (2006) Principal Components Analysis corrects for Stratification in Genome-Wide Association Studies [7].

Chromosome	High-quality markers	% Random markers
1	43,959	7,832
2	46,464	8,278
3	38,737	6,901
4	34,545	6,154
5	35,058	6,246
6	39,287	6,999
7	31,266	5,570
8	31,809	5,667
9	27,083	4,825
10	29,943	5,334
11	27,893	4,969
12	27,746	4,943
13	21,525	3,835
14	18,773	3,345
15	17,045	3,037
16	17,129	3,052
17	14,926	2,659
18	16,914	3,013
19	9,964	1,775
20	14,307	2,549
21	8,349	1,487
22	8,587	1,530
Total	**561,309**	**100,000**

Table 4.3: Number of randomly chosen high-quality markers by chromosome.



(a) PC1 and PC2 of cases group. Red: European-American individuals.



(b) PC1 and PC2 of controls group. Red: European-American individuals.

Figure 4.3: Principal Component Analysis of the samples, based on all ethnicities.

Sample Relatedness

The data used in this research correspond to a family study in which there were up to two relatives per affected individual. If subjects are treated as independent samples, the results can bias the study increasing false positives and false negatives [49]. For this reason, only one member per family was randomly selected, with the information provided in the pedigree data *(pht000709.v2.p2)* [15]. This selection generated a sample of 1,830 unrelated European-American subjects (n cases = 813, n controls = 1,017) denoted by **CCun**. An experiment was run with another randomly chosen family member to verify that the random selection did not affect the model building. Since both models had similar performances, the construction of the models is not dependent on the selected family member, which means that a random selection did not bias the models.

4.2.2 Marker Quality

The original dataset has 620,901 SNPs genotyped by the Center for Inherited Disease Research (CIDR) using the Illumina Infinium II assay protocol with hybridization to Illumina Human 610Quadv1_B Beadchips. This research only takes into account the autosomal chromosomes, so the markers of chromosome 23 were removed, and 600,470 SNPs remained. The quality control procedures of the genotypes were carried out separately in cases and controls and, the union of both sets of SNPs gave rise to the set of 561,039 high-quality markers. Table 4.4 shows the statistics of the quality control procedures performed on the case and control genotypes. Appendix 9.1 presents a more detailed summary, where the data is displayed by chromosome.

The first filter consisted of removing those markers that did not have a reference SNP ID number, or "rs" ID, which is an identification tag assigned by NCBI to a group of SNPs that map to an identical location. Due to this, 2.99% of the initial markers were eliminated, leaving a set of 582,539 SNPs.

Call Rate

Genotyping efficiency or call rate is the fraction of called markers per sample over the total number of SNPs in the dataset [49]. Markers with a low call rate should be filtered out since they indicate poor assays and cause spurious data. The cut-off criterion was to remove those SNPs with less than 98% call rate as suggested in [17]. This procedure removed 2.38% and 2.50% of the markers in cases and controls respectively.

Monomorphic markers

Monomorphic markers are those that have the same genotype for all individuals. They are uninformative since they do not reveal a genotypic difference. This procedure filtered out

1.85% of the markers in cases and 2.04% in controls.

Hardy-Weinberg Equilibrium

The Hardy-Weinberg Equilibrium (HWE) principle states that genetic variation in a large, randomly mating population will remain constant from one generation to the next as long as there are no evolutionary influences. The concepts of HWE have been used to detect genotyping errors [25]. This filtering procedure applies only to controls since in cases the deviations of the HWE could show association with the disease. The threshold of significance to defining which SNPs are in HWE varies between studies, some p-values cut-off criteria previously reported in other studies range from 0.001 to 5×10^{-8} [10, Chapter 7]. In this research the SNPs with a p-value less than 1×10^{-5} (0.17% of the markers) were removed as suggested in [17]. As previously mentioned, the union of the filtered sets of cases and controls originated the 561,309 high-quality markers set that was used for the pre-filtering stage presented in the next chapter.

	Criterion	Before QC	Number of Markers Removed	After QC	% of Markers Removed
Cases	Markers without reference SNP ID number (“rs” ID)	600,470	17,931	582,539	2.99
	Call Rate < 98%	582,539	13,857	568,682	2.38
	Monomorphic markers	568,682	10,796	557,886	1.85
Controls	Markers without reference SNP ID number (“rs” ID)	600,470	17,931	582,539	2.99
	Call Rate < 98%	582,539	14,588	567,951	2.50
	Monomorphic markers	567,951	11,910	556,041	2.04
	Hardy-Weinberg Equilibrium (p <1 × 10^{-5})	556,041	977	555,064	0.17
Total	**Union between cases and controls**	**600,470**	**39,431**	**561,309**	**6.57**

Table 4.4: Results of the quality control procedures on genotype data of cases and controls.

4.3 Summary

This chapter presented the data used in this study and the pre-processing that was carried out. The importance of quality control procedures in samples and markers in association studies was also mentioned. Moreover, the chapter details the quality filters used in this research to reduce false positives and false negatives. The results presented in this section generated the high-quality marker dataset that was used in the pre-filtering stage shown in the next chapter.

5 | Dimensionality Reduction by Feature Pre-selection

5.1 Univariate Method

The quality control procedures originated a dataset of 561,309 high-quality SNPs. It was decided to pre-filter with a univariate method to reduce the number of features including only the most significant markers in the association of AD. As previously mentioned, this strategy had two objectives: (1) to take advantage of the computational efficiency of the univariate methods and (2) to replicate the study presented in [17] to verify that the data were handled correctly through the pre-processing and quality control stages.

5.1.1 Corrected Chi-squared Test

As mentioned in chapter 1, chi-squared tests are the most commonly used univariate methods to evaluate the association between markers and phenotypes when comparing allele frequencies between cases and controls. However, GWAS can be affected by increased false positive results in the presence of unrecognized cryptic relatedness between subjects. Since the data used in this research comes from a family study, an approach suggested by [56] was taken to correct for relatedness by estimating kinship coefficients that is, a corrected χ^2 test.
The relationship between individuals can be measured through the kinship matrix denoted by:

$$\Phi = 2 \begin{bmatrix} \phi_{1,1} & \cdots & \phi_{1,N} \\ \vdots & \phi_{i,j} & \vdots \\ \phi_{N,1} & \cdots & \phi_{N,N} \end{bmatrix} \tag{5.1}$$

where $\phi_{i,j}$ is the kinship coefficient between subject i and j.

The statistical test equivalent to the $\chi^2_{corrected}$ test can be calculated based on the minor allele frequencies (MAFs) estimated in affected individuals (p_{aj}) and all subjects (p_j) [38] by:

$$W_{corrected} = \frac{4N_a^2\left(\sum_{j=1}^m p_{aj} - \sum_{j=1}^m p_j\right)^2}{cs\left(Y'\Phi Y - \frac{2N_a}{N}1'_N\Phi Y + \frac{N_a^2}{N^2}1'_N\Phi 1_N\right)} \tag{5.2}$$

where N_a is the number of affected subjects, $N = N_a + N_u$ is the number of all subjects (affected and unaffected), Y is a phenotype vector of length N with i^{th} entry $Y_i = 1$ if the subject is a case and $Y_i = 0$ if is a control, 1_N is a vector of length N with all entries equal to 1, Φ is the kinship matrix and cs is the value of the correlation between SNPs computed as:

$$cs = \sum_{k=1}^{m}\sum_{j=1}^{m} v_{kj} \tag{5.3}$$

where $v_{kj} = cov(G_k^w, G_j^w)$ if $k \neq j$ and $v_{jj} = w_j var(G_j) = 2w_j p_j(1-p_j)$ if $k = j$.
G is the genotype matrix of m SNPs, where $g_{i,j}$ is the genotype of the i^{th} subject at the j^{th} SNP coded as 0, 1 or 2 copies of the minor allele.

$$G = \begin{bmatrix} G_1 & \cdots & G_j & \cdots & G_m \end{bmatrix} = \begin{bmatrix} g_{1,1} & \cdots & g_{1,m} \\ \vdots & g_{i,j} & \vdots \\ g_{N,1} & \cdots & g_{N,m} \end{bmatrix} \tag{5.4}$$

And G^w is the weighted genotype matrix.

$$G = \begin{bmatrix} G_1^w & \cdots & G_j^w & \cdots & G_m^w \end{bmatrix} = \begin{bmatrix} g_{1,1}^w & \cdots & g_{1,m}^w \\ \vdots & g_{i,j}^w & \vdots \\ g_{N,1}^w & \cdots & g_{N,m}^w \end{bmatrix} \tag{5.5}$$

where $g_{i,j}^w = \sqrt{w_j} g_{i,j}$ and w_j is a weight for SNP j computed as $w_j = \frac{1}{\sqrt{p_{uj}(1-p_{uj})}}$.

In summary, to calculate the $\chi^2_{corrected}$ test, the following data are required:

1. MAF matrix *(#SNPs x 2)*, where the first column corresponds to cases and the second one to controls.

2. Phenotype matrix *(#subjects x 1)*, where 0's represent controls and 1's cases.

3. Kinship matrix *(#subjects x #subjects)* computed with genotype and pedigree data.

4. Correlation matrix *(#SNPs x #SNPs)*

Positive Control

A "positive control" experiment was conducted to assess the validity of the χ^2 test performed in this study. The experiment consisted of generating a database of 344 SNPs previously reported in [17] and 200 random SNPs from the high-quality markers set. The objective was to verify that the p-values reported were similar to those calculated in this study, to validate the handling of the data and the quality control procedures. The results proved a high positive correlation $(R^2 = 0.99)$ as shown in figure 5.1. Therefore, the replication of the study presented in the literature was successful.

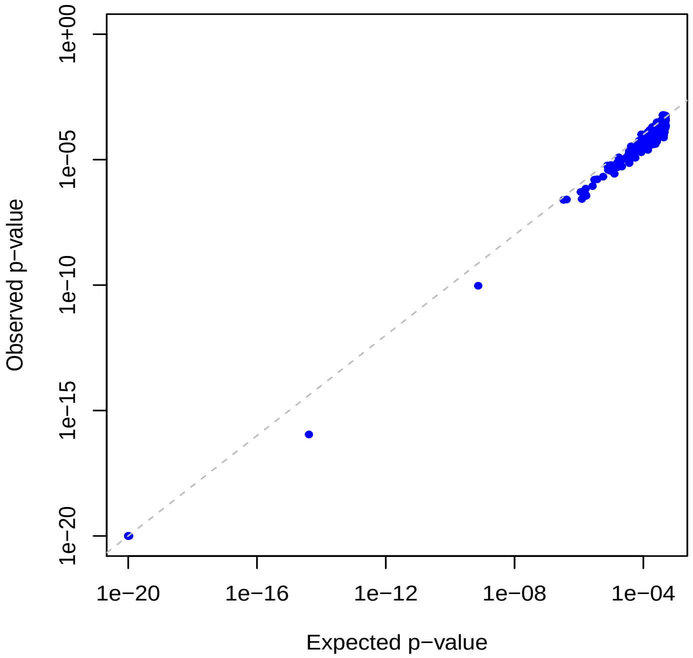


Figure 5.1: Positive control results: Expected p-values vs Observed p-values ($log10$ scale), $R^2 = 0.99$.

The χ^2 test was performed with a dataset that included the individuals of CCall and the 561,309 high-quality markers. As expected, SNPs close to APOE on chromosome 19 show a stronger association with AD. The obtained p-values are displayed in the Manhattan plot of figure 5.2.

Once the p-values of the markers were obtained, those with a p-value lower than 1×10^{-3} were selected. Figure 5.3 shows the distribution of all the p-values of the SNPs. This preselection stage originated a set of 1,106 markers which composed the dataset denoted by **Th-CCun** *(1,106 SNPs x 1,830 subjects)*. Furthermore, the **Top100-CCun** *(100 SNPs x 1,830 subjects)* dataset was created, which includes the top 100 markers and the individuals of the CCun group.

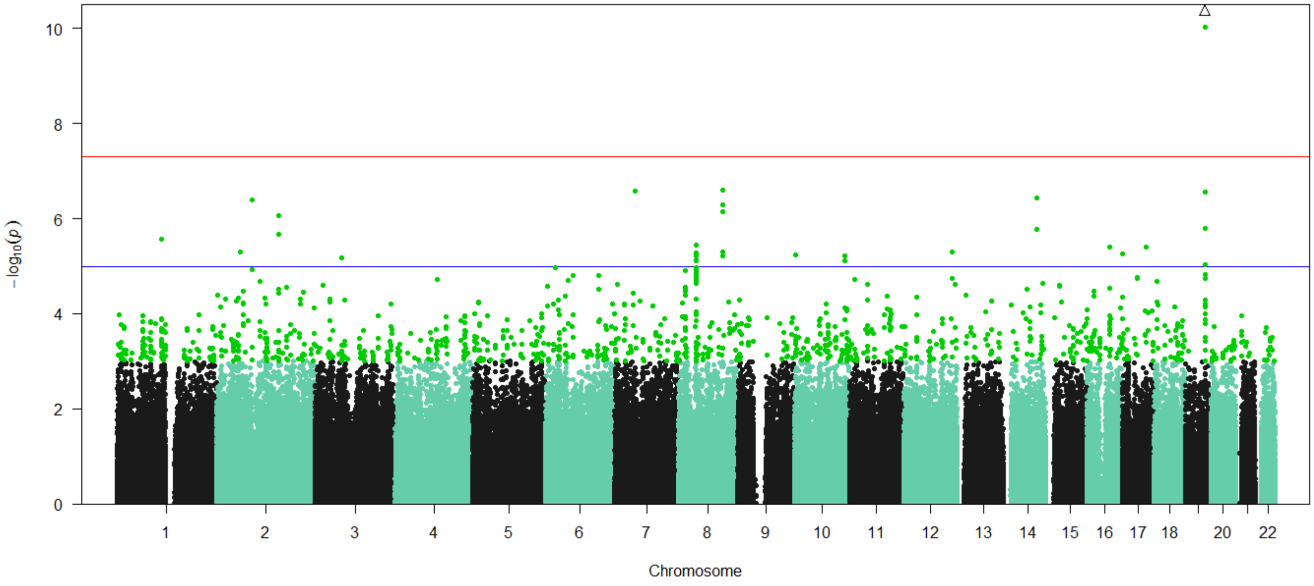


Figure 5.2: Manhattan plot for GWAS analysis of AD based on CCall sample. Green dots: SNPs with p-value $< 1 \times 10^{-3}$, red line: genome-wide significance level $-log10(5 \times 10^{-8})$, blue line: suggestive line $-log10(1 \times 10^{-5})$. Plots have been truncated at $-log10(p) = 10$ on the vertical axis to more clearly visualize the results for most of the genome. Some markers near APOE on chromosome 19 yielded $-log10(p) \gg 10$, and are represented by a triangle at the top of the panel.

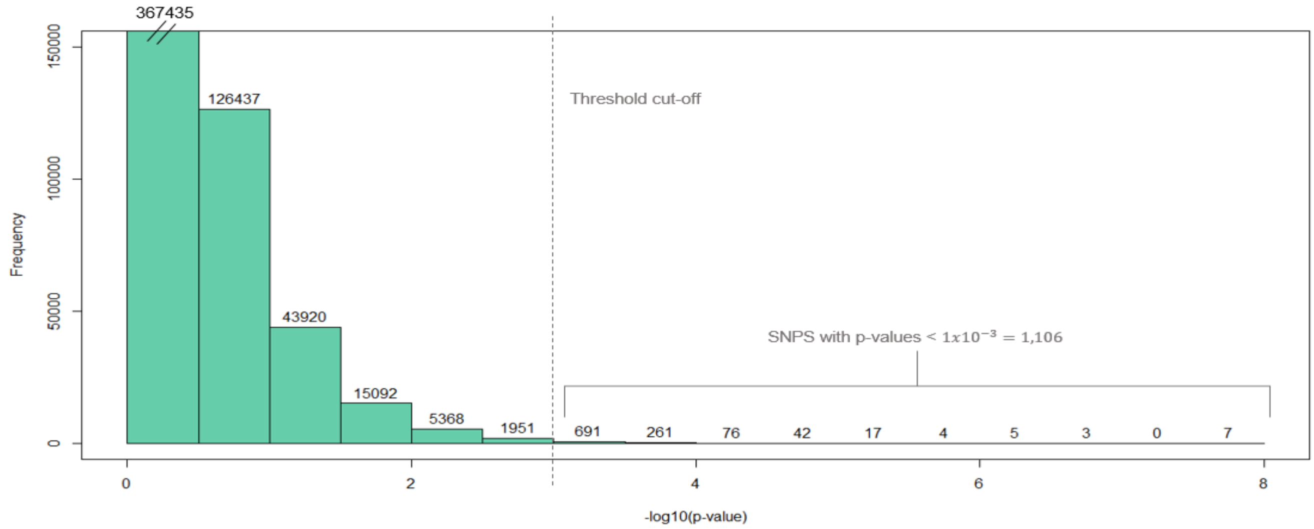


Figure 5.3: Distribution of the p-values. The selection threshold used in this study was to select markers with p-values <1×10^{-3}, which originated a set of 1,106 SNPs. The frequency on the vertical axis was limited to 150,000 to emphasize the markers with the most significant p-values.

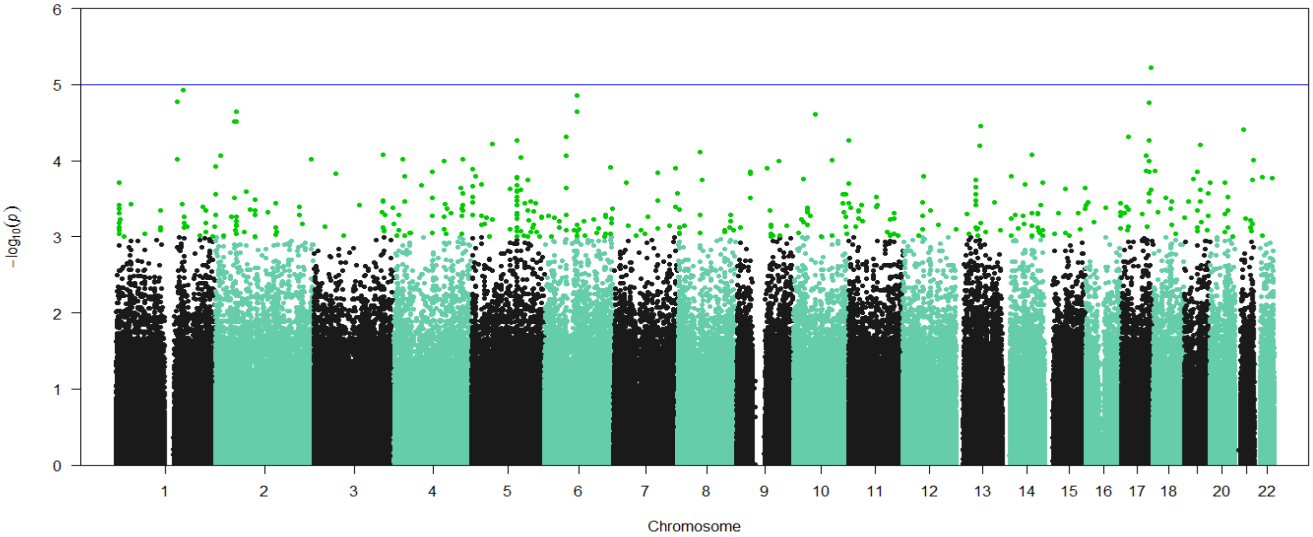


Figure 5.4: Manhattan plot for GWAS analysis of AD based on misclassified samples. Green dots: SNPs with p-value $< 1 \times 10^{-3}$, blue line: suggestive line $-log10(1 \times 10^{-5})$.

For the analysis of the misclassified samples, a chi-squared test with kinship correction was also performed to identify the variants with higher association within the group of the poorly classified samples. Figure 5.4 shows the Manhattan plot of the markers while figure 5.5 shows the distribution of the p-values. For this experiment, the same selection threshold of 1×10^{-3} was taken and originated a set of 461 highly associated markers.

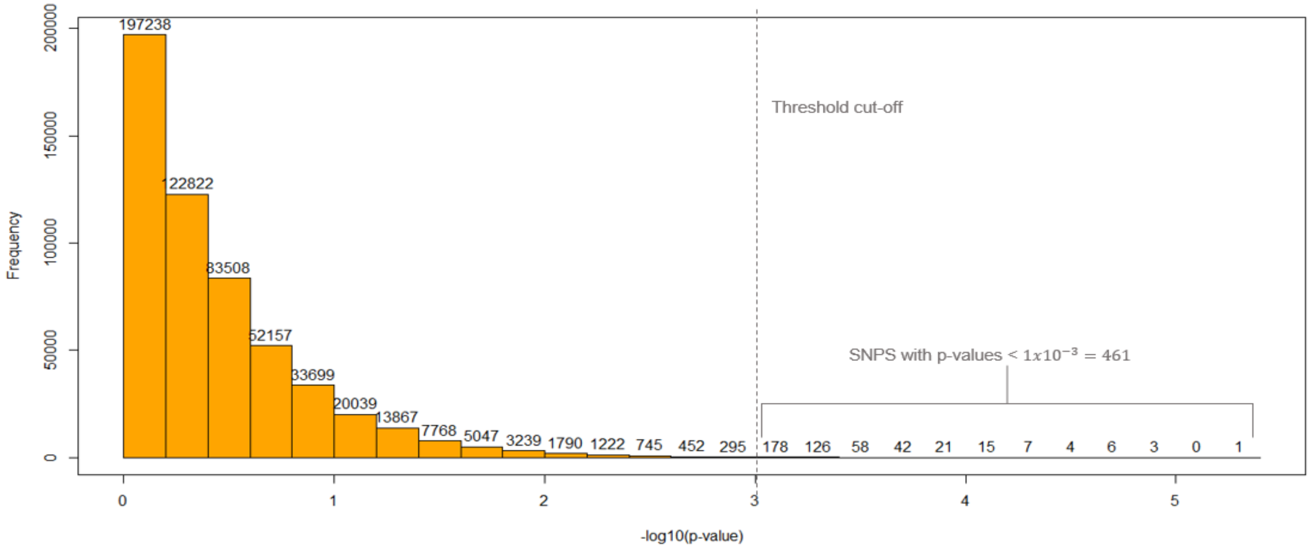


Figure 5.5: Distribution of the p-values. The selection threshold used in this study was to select markers with p-values <1×10^{-3}, which originated a set of 461 SNPs.

5.2 Summary

This chapter explained the univariate approach used to pre-select markers to reduce dimensionality. Moreover, section 5.1.1 describes a validation experiment known as a positive control, which aimed to evaluate the handling of the data through the replication of a study previously presented in the literature. From 561,309 SNPs, those with p-value $< 1 \times 10^{-3}$ were filtered, generating a set of 1,106 SNPs. The results were shown in a Manhattan plot. Finally, two pre-selected datasets were described: Th-CCun and Top100-CCun. Both datasets are used as input data in the machine learning methods presented in the next chapter.

6 | Model Building

This study proposes the use of machine learning methods to build predictive models that maximize discrimination between healthy individuals and individuals with AD based on GWAS data. The advantages of ML methods against traditional techniques are their robustness and their ability to consider interactions between features. This chapter details the two approaches that were chosen to build the models: wrapper and embedded methods. Within the wrapper methods, two algorithms were tested: BSWiMS and GA + SVM, while LASSO was the embedded method used. All algorithms were tested with two datasets: Top100-CCun (top 100 high-quality SNPs and 1,830 unrelated individuals) and Th-CCun (1,106 high-quality SNPs and 1,830 unrelated individuals). The missing values of both datasets (0.078% of the top100-CCun and 0.093% of the Th-CCun) were imputed by assigning the median values of the nearest neighbors with the function *nearestNeighborImpute()* of FRESA.CAD [5].

6.1 Wrapper Methods

As mentioned in chapter 2, wrapper methods have three main characteristics: (1) they are based on a search algorithm that examines the entire solution space, (2) they have a function that evaluates the prediction accuracy of the subsets of features and, (3) they incorporate a learning algorithm to the feature selection stage. According to the search strategy, wrapper methods can be classified into two categories: greedy search algorithms and heuristic search algorithms.

6.1.1 Greedy Search Algorithm

BSWiMS: Bootstrapping Stage-Wise Model Selection

BSWiMS is a supervised method based on statistics that generates a logistic regression model to test variables and build larger models. The algorithm consists of the following phases:

1. Univariate Filter: Estimates the univariate association of each SNP with the phenotype and the association p-value is FDR adjusted using the Benjamini–Hochberg procedure.

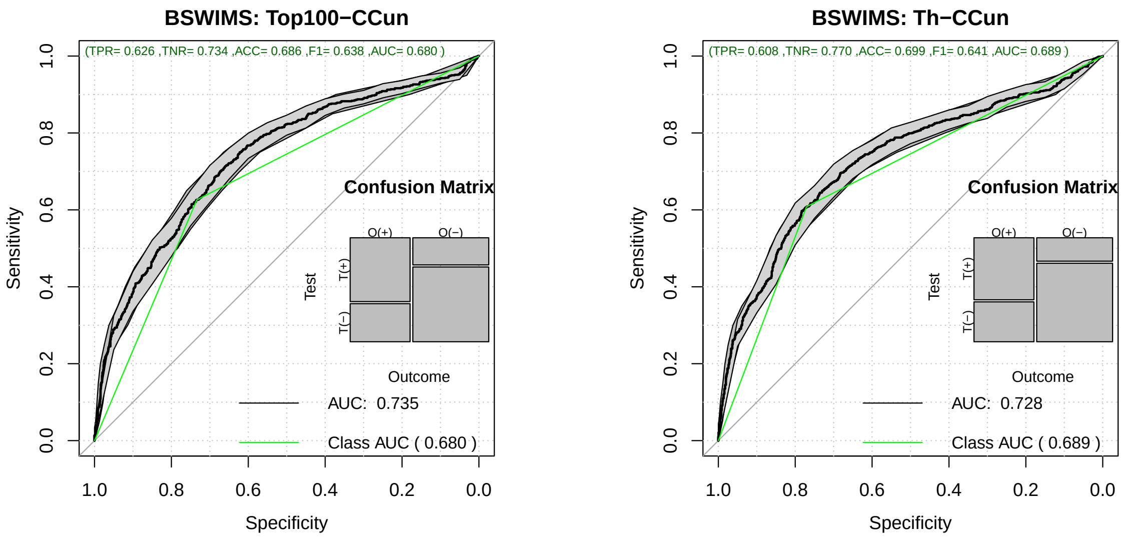
2. Bootstrapped Forward Selection: It builds linear models using FS by adding features that increase the classification.

3. Frequency-based Forward Selection: The markers of each of the bootstrapped models are ranked according to their selection frequency. A single model is generated via FS step-wise by adding the ranked SNPs if and only if they increase the model classification.

4. Bootstrapped Backwards Elimination: It removes one marker at a time from the forward model and bootstraps the reduced model. If the reduced model delivers the same information as the entire model, then the variable is eliminated by creating a compact model. Every time a compact model is found, its features are removed and the process repeats steps 2-4 until there are no new compact models.

5. Model Bagging: Combine all compact models in a single statistical model.

As mentioned above, the datasets Top100-CCun and Th-CCun were used to build the model that maximizes discrimination between cases and controls. Steps 1-4 were repeated 25 times to infer the importance of each SNP in the prediction. Subsequently, the performance of the model was evaluated 20 times by CV, splitting the dataset in 80% for training and 20% for testing. Figure 6.1 displays the ROC curves and the confusion matrices generated by the models of both datasets.



(a) Model with Top100-CCun dataset.

(b) Model with Th-CCun dataset.

Figure 6.1: ROC curves and confusion matrices for BSWIMS models. Black line: continuous prediction, green line: operation point.

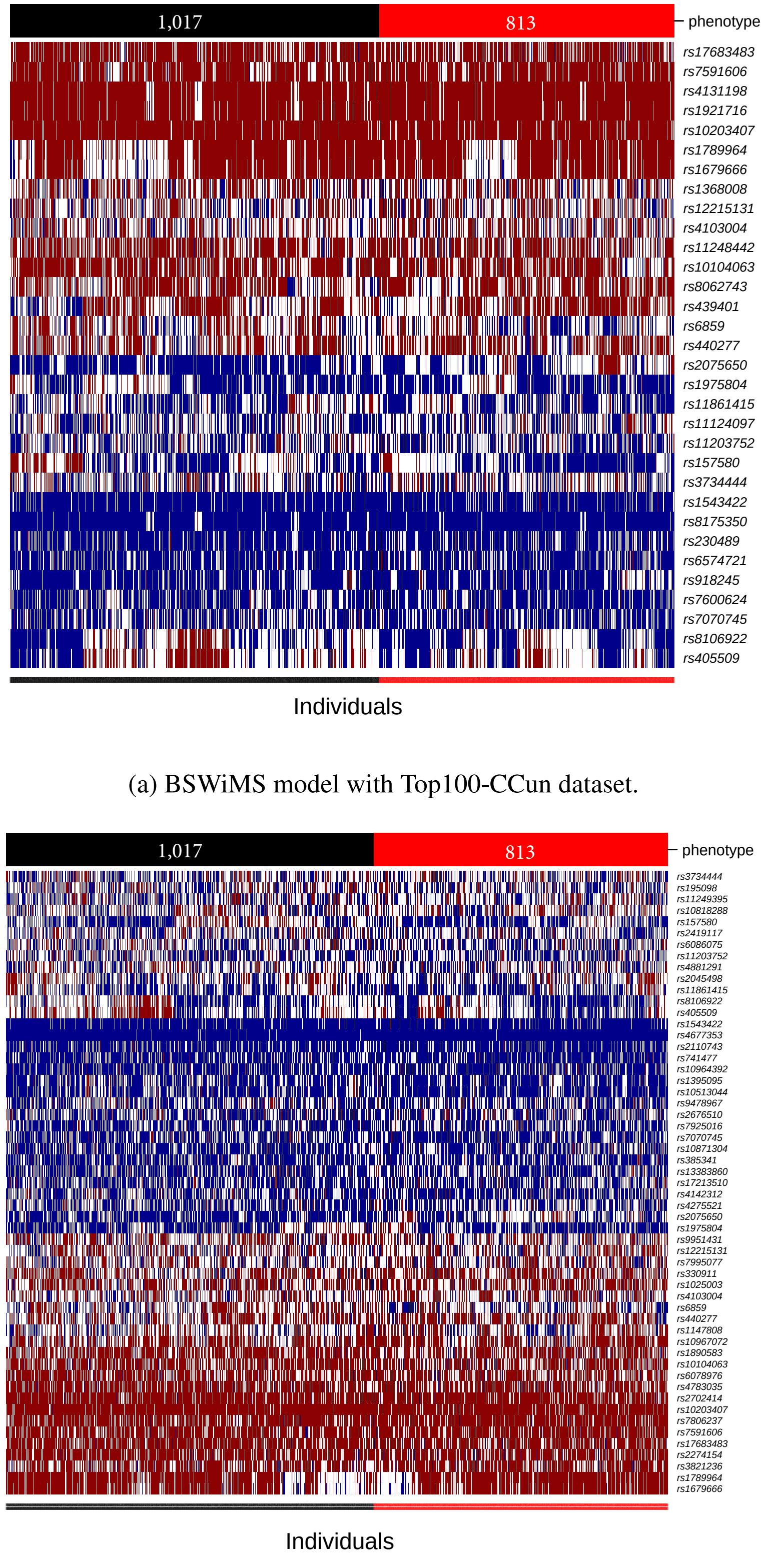
Tables 6.1 and 6.2 show the SNPs of each model with their respective coefficients and each variable contribution to the model in terms of accuracy and Area Under the Curve. Table 6.3 summarizes the results of the evaluation of the performance of the models with different metrics. The model built with the Top100-CCun dataset obtained an accuracy of 0.686, a sensitivity of 0.626, a specificity of 0.733 and an AUC of 0.68. On the other hand, the Th-CCun model had a slightly better performance with an accuracy of 0.699, a specificity of 0.77 and an AUC of 0.689; however, sensitivity decreased to 0.608. Figure 6.2 shows a way to visualize the markers of the models and the genotypes of the individuals with a heatmap, where the allele frequencies are represented with 0,1 or 2. The objective of these plots is to visualize the differences in the frequency of alleles of cases and controls for each model SNP.

SNP	Estimate	Accuracy			AUC			Frequency
		Univariate	Removing	Multivariate	Univariate	Removing	Multivariate	
rs2075650	0.5248	0.6847	0.6237	0.7002	0.6791	0.6024	0.6909	1
rs157580	-0.6668	0.5913	0.6796	0.7075	0.5898	0.6680	0.6982	1
rs405509	-0.4005	0.5940	0.7076	0.7075	0.5725	0.6987	0.6982	1
rs8106922	-0.3663	0.5814	0.7038	0.7075	0.5757	0.6939	0.6982	1
rs10203407	-0.3376	0.5727	0.6986	0.7002	0.5244	0.6886	0.6909	1
rs11861415	-0.1161	0.5557	0.7015	0.7002	0.5000	0.6931	0.6909	1
rs7070745	-0.1340	0.5557	0.7082	0.7002	0.5000	0.6997	0.6909	1
rs17683483	-0.1417	0.5617	0.6956	0.7002	0.5332	0.6861	0.6909	1
rs1543422	0.2235	0.5705	0.6939	0.7002	0.5257	0.6840	0.6909	1
rs7591606	0.1607	0.5557	0.7054	0.7002	0.5000	0.6964	0.6909	1
rs10104063	-0.1557	0.5634	0.7015	0.7032	0.5443	0.6924	0.6939	1
rs3734444	0.1020	0.5557	0.6979	0.7002	0.5206	0.6887	0.6909	1
rs11203752	-0.1675	0.5557	0.7046	0.7075	0.5000	0.6956	0.6982	1
rs6859	-0.2933	0.5978	0.7017	0.7085	0.5677	0.6924	0.6991	0.92
rs440277	0.2566	0.5557	0.7042	0.7087	0.5000	0.6949	0.6993	0.88
rs439401	0.2055	0.5585	0.7032	0.7087	0.5558	0.6944	0.6993	0.88
rs11124097	0.1620	0.5585	0.7073	0.7079	0.5154	0.6978	0.6986	0.88
rs8062743	0.1747	0.5557	0.7053	0.7091	0.5000	0.6968	0.6996	0.8
rs4103004	0.1529	0.5557	0.7081	0.7085	0.5000	0.6997	0.6990	0.8
rs1975804	-0.1428	0.5557	0.7086	0.7082	0.5000	0.6993	0.6988	0.68
rs918245	0.1230	0.5656	0.7043	0.7055	0.5381	0.6952	0.6960	0.52
rs4131198	0.2077	0.5557	0.7066	0.7082	0.5000	0.6971	0.6987	0.44
rs7600624	-0.1337	0.5557	0.7128	0.7114	0.5000	0.7037	0.7019	0.44
rs8175350	-0.1503	0.5557	0.7093	0.7102	0.5000	0.6998	0.7008	0.32
rs1368008	-0.0344	0.5557	0.7038	0.7093	0.5000	0.6945	0.6997	0.28
rs1679666	0.0542	0.5557	0.7088	0.7080	0.5000	0.6994	0.6985	0.24
rs12215131	0.0394	0.5557	0.7062	0.7098	0.5000	0.6969	0.7008	0.24
rs230489	0.0683	0.5667	0.7121	0.7146	0.5354	0.7028	0.7055	0.2
rs6574721	0.0331	0.5590	0.7077	0.7082	0.5310	0.6985	0.6984	0.16
rs11248442	-0.0380	0.5579	0.7143	0.7150	0.5079	0.7050	0.7058	0.16
rs1921716	0.0347	0.5557	0.6998	0.7022	0.5000	0.6915	0.6934	0.12
rs1789964	0.0114	0.5557	0.6943	0.7005	0.5000	0.6864	0.6923	0.08

Table 6.1: Summary of the 32 SNPs of the BSWiMS model generated with Top100-CCun dataset. The contribution of each variable to the model is expressed in terms of accuracy and Area Under the Curve. **Univariate**: the value of acc/AUC of the model including only that variable, **Removing**: the value of acc/AUC of the model when removing that variable and **Multivariate**: the acc/AUC of the model when including that variable.

SNP	Estimate	Accuracy			AUC			Frequency
		Univariate	Removing	Multivariate	Univariate	Removing	Multivariate	
rs2075650	1.0100	0.6842	0.6530	0.7298	0.6786	0.6399	0.7210	1
rs6078976	-0.2938	0.5590	0.7231	0.7298	0.5067	0.7140	0.7210	1
rs11861415	-0.2295	0.5557	0.7257	0.7298	0.5000	0.7169	0.7210	1
rs4677353	-2.2344	0.5557	0.7276	0.7298	0.5000	0.7185	0.7210	1
rs1543422	0.4614	0.5705	0.7241	0.7298	0.5257	0.7151	0.7210	1
rs13383860	0.2682	0.5623	0.7274	0.7293	0.5118	0.7184	0.7206	1
rs7070745	-0.2648	0.5557	0.7265	0.7298	0.5000	0.7175	0.7210	1
rs2110743	-0.3162	0.5557	0.7252	0.7303	0.5000	0.7165	0.7216	0.96
rs385341	0.2737	0.5617	0.7258	0.7298	0.5092	0.7167	0.7210	0.96
rs7806237	0.2989	0.5557	0.7222	0.7278	0.5000	0.7134	0.7192	0.92
rs3821236	0.1899	0.5557	0.7205	0.7253	0.5000	0.7114	0.7166	0.88
rs10203407	-0.4987	0.5727	0.7255	0.7303	0.5244	0.7161	0.7215	0.88
rs1890583	-0.2405	0.5623	0.7270	0.7302	0.5100	0.7179	0.7214	0.8
rs4142312	-0.1773	0.5557	0.7249	0.7286	0.5000	0.7160	0.7197	0.8
rs2274154	-0.1813	0.5623	0.7186	0.7236	0.5328	0.7098	0.7148	0.72
rs157580	-0.4699	0.5907	0.6433	0.7189	0.5892	0.6303	0.7101	0.68
rs405509	-0.2322	0.5940	0.7083	0.7189	0.5725	0.6992	0.7101	0.68
rs8106922	-0.2181	0.5809	0.7135	0.7189	0.5750	0.7043	0.7101	0.68
rs10871304	0.1405	0.5617	0.7166	0.7221	0.5101	0.7074	0.7135	0.68
rs10967072	0.1453	0.5557	0.7260	0.7301	0.5000	0.7170	0.7214	0.64
rs330911	0.1470	0.5557	0.7250	0.7285	0.5000	0.7164	0.7195	0.6
rs9951431	-0.0790	0.5541	0.7147	0.7195	0.5138	0.7057	0.7109	0.52
rs6086075	-0.0603	0.5557	0.7089	0.7168	0.5000	0.6999	0.7077	0.48
rs9478967	-0.0712	0.5557	0.7149	0.7209	0.5000	0.7060	0.7122	0.44
rs17683483	-0.1234	0.5617	0.7220	0.7276	0.5332	0.7128	0.7185	0.44
rs7925016	0.0735	0.5612	0.7145	0.7197	0.5096	0.7054	0.7108	0.4
rs11203752	-0.0577	0.5557	0.7132	0.7181	0.5000	0.7044	0.7094	0.4
rs1147808	0.0692	0.5557	0.7248	0.7268	0.5000	0.7157	0.7176	0.36
rs1679666	0.0609	0.5557	0.7109	0.7177	0.5000	0.7013	0.7094	0.32
rs2045498	-0.0469	0.5557	0.7115	0.7199	0.5000	0.7029	0.7112	0.32
rs7995077	0.0503	0.5749	0.7186	0.7207	0.5536	0.7096	0.7120	0.32
rs4783035	-0.0550	0.5574	0.7102	0.7164	0.5047	0.7011	0.7077	0.32
rs4275521	0.0714	0.5579	0.7293	0.7346	0.5091	0.7207	0.7258	0.32
rs2676510	0.0436	0.5617	0.7137	0.7187	0.5154	0.7050	0.7097	0.28
rs7591606	0.0548	0.5557	0.7079	0.7139	0.5000	0.6986	0.7046	0.24
rs4103004	0.0328	0.5557	0.7068	0.7142	0.5000	0.6972	0.7051	0.24
rs1975804	-0.0261	0.5557	0.7052	0.7122	0.5000	0.6956	0.7026	0.16
rs4881291	-0.0290	0.5557	0.7247	0.7273	0.5000	0.7162	0.7191	0.16
rs10964392	0.0265	0.5601	0.7014	0.7049	0.5347	0.6912	0.6949	0.16
rs6859	-0.0324	0.5978	0.6960	0.7122	0.5677	0.6856	0.7025	0.12
rs440277	0.0298	0.5557	0.7016	0.7122	0.5000	0.6923	0.7025	0.12
rs1789964	0.0295	0.5557	0.7106	0.7208	0.5000	0.7011	0.7115	0.12
rs2419117	-0.0285	0.5557	0.7262	0.7266	0.5000	0.7171	0.7174	0.12
rs10513044	-0.0220	0.5557	0.7056	0.7051	0.5000	0.6952	0.6956	0.12
rs10818288	0.0242	0.5623	0.7275	0.7342	0.5340	0.7189	0.7252	0.12
rs10104063	-0.0100	0.5634	0.6874	0.6932	0.5443	0.6773	0.6821	0.08
rs195098	-0.0114	0.5557	0.7150	0.7167	0.5000	0.7052	0.7078	0.08
rs11249395	0.0111	0.5508	0.7093	0.7131	0.5140	0.7000	0.7042	0.08
rs1395095	-0.0259	0.5557	0.7322	0.7374	0.5000	0.7241	0.7292	0.08
rs741477	-0.0230	0.5557	0.7287	0.7352	0.5000	0.7191	0.7263	0.08
rs2702414	-0.0182	0.5683	0.6954	0.6970	0.5301	0.6840	0.6859	0.08
rs3734444	0.0159	0.5557	0.7270	0.7339	0.5206	0.7172	0.7247	0.08
rs12215131	0.0064	0.5557	0.6962	0.6959	0.5000	0.6857	0.6858	0.08
rs17213510	0.0196	0.5617	0.7276	0.7377	0.5117	0.7188	0.7292	0.08
rs1025003	0.0115	0.5557	0.6951	0.6970	0.5000	0.6851	0.6859	0.08

Table 6.2: Summary of the 55 SNPs of the BSWiMS model generated with Th-CCun dataset.The contribution of each variable to the model is expressed in terms of accuracy and Area Under the Curve.**Univariate**: the value of acc/AUC of the model including only that variable,**Removing**: the value of acc/AUC of the model when removing that variable and **Multivariate**: the acc/AUC of the model when including that variable.



(a) BSWiMS model with Top100-CCun dataset.

(b) BSWiMS model with Th-CCun dataset.

Figure 6.2: Heatmaps of the markers selected by each model and individual genotype data (blue:0, white:1, red:2).

6.1.2 Heuristic Search Algorithm

Genetic Algorithms

The heuristic approach for the model building uses an R package based on genetic algorithms coupled to statistical modeling methods for supervised classification called GALGO [53]. The process to generate the model consists of the following steps:

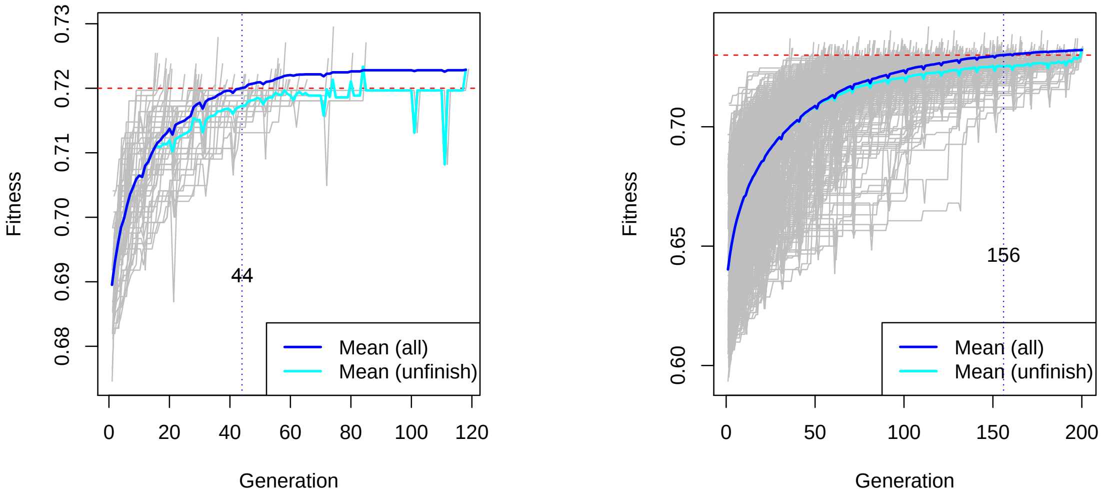
1. Setting-up the analysis
   This stage includes (1) the specification of the datasets with the markers per row and samples per column (Top100-CCun and Th-CCun), (2) the definition of the parameters for the GA search, including 50 solutions per Top100-CCun model and 1,000 solutions per Th-CCun model, and a subset size of 25 features evaluated in each solution, (3) the election of SVM with radial kernel as a classification method and, (4) the error estimation by 3-folds CV.

2. Search for multivariate models
   The genetic algorithm explores the feature space to find combinations of markers that maximize class discrimination using SVM and generates 50 solutions.

3. Analysis of chromosomes
   The classification accuracy of the chromosomes (subsets of 25 features generated in the previous stage) is evaluated by 3-fold CV. Figure 6.3 (a) shows that on average a solution is reached in generation 44 for the Top100-CCun dataset and (b) in generation 156 for Th-CCun dataset. The convergence of the model is assessed analyzing the markers frequency in the chromosome population.

   Figure 6.4 displays the 50 most frequent SNPs ordered by rank on the horizontal axis and by frequency on the vertical axis. The marker rs2075650 has the highest frequency in both experiments and has been stable during the (a) 50 and (b) 1000 solutions. The SNPs of the Top100-CCun dataset tend to stabilize faster than those of the Th-CCun. When performing a stochastic search, the 100 markers are more likely to appear on the chromosomes since the search space is smaller and stabilized in 50 solutions, unlike the set of 1,106 SNPs which requires more solutions to be stable.

4. Development of a representative model
   This stage builds a model through a forward selection procedure including the most frequent SNPs in the population of chromosomes generated. Figure 6.5 shows the graphs of the models using FS. The results showed a slight difference between the models generated with the different datasets, as shown in table 6.3. For the solutions of Top-CCun, the FS procedure built 14 models where model labeled as 13 with 49 features was the best with a maximum accuracy of 0.72, a sensitivity of 0.616, a specificity of 0.8 and an AUC of 0.708. On the contrary, for Th-CCun dataset 13 models were generated by FS where model identified as 5 with 42 markers was the best with an accuracy of 0.7264, a sensitivity of 0.626, a specificity of 0.806 and an AUC of 0.716. The confusion matrices shown in figure 6.6 are another way to represent the results of the classification.



(a) Fitness evolution with Top100-CCun dataset in 50 independent searches.

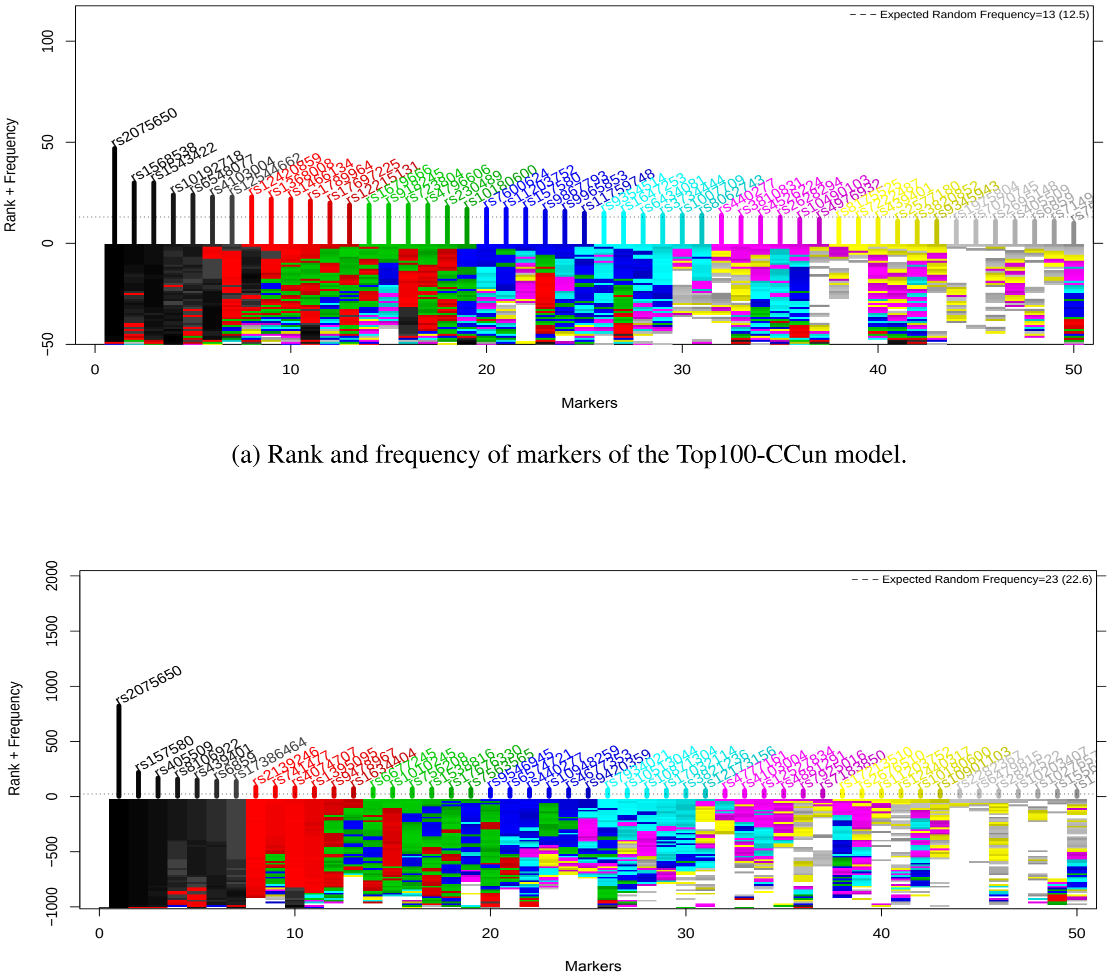
(b) Fitness evolution with Th-CCun dataset in 1,000 independent searches.

Figure 6.3: Evolution of the maximum fitness across generations.

There is a similarity in the results of both models where the FN almost duplicate the FP. As subsequent experiments to this research, it is suggested to examine the misclassified markers to analyze the possible reasons that lead to a poor classification.

6.2 Embedded Method

LASSO is one of the best-known embedded methods and uses L1-regularization as a penalty against complexity to reduce the degree of overfitting of a model. The penalty term introduces sparsity by performing a type of feature selection as part of the training of the model. LASSO built the model with the same sets as the BSWiMS method (80% for training and 20% for testing). The performance of the model was evaluated through cross-validation. The lambda parameter was estimated using 10-fold CV to obtain the minimum mean cross-validated error. The two models generated with LASSO had a better performance than those built with the wrapper methods as shown in table 6.3. The model trained on Top100-CCun dataset consisted of 71 features and had an accuracy of 0.7656, a sensitivity of 0.6626, a specificity of 0.8246 and an AUC of 0.7436. On the other hand, the model of 433 markers generated with Th-CCun outperformed all models with an accuracy of 0.8013, a sensitivity of 0.7975, a specificity of 0.8035 and an AUC of 0.8005, figure 6.7 shows the ROC curves of each model. These results are promising since this study uses only genetic data and Alzheimer's disease is a complex disease that is believed to include a combination of genetic, environmental, and lifestyle factors. A list with the SNPs of each model is shown in Appendix B 9.2.



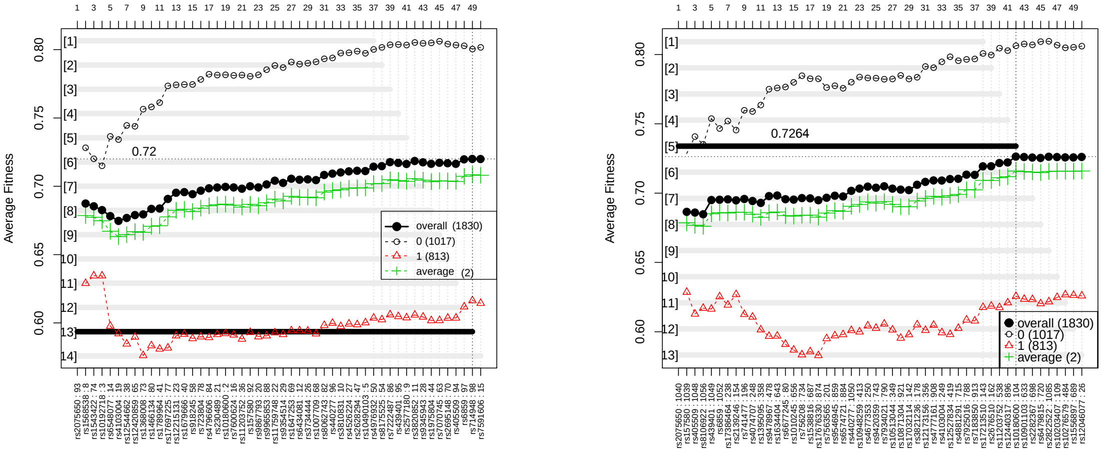
(a) Rank and frequency of markers of the Top100-CCun model.

(b) Rank and frequency of markers of the Th-CCun model.

Figure 6.4: Rank stability across past evolutions.

6.2.1 Misclassified Re-Modeling and Mixture Models

This stage included a detailed analysis of the poorly classified samples of the LASSO model with Th-CCun corresponding to 19.86% of the total samples. The chi-squared test with kinship correction was performed again in the subset of 89 poorly classified individuals. As mentioned in section 5.1.1, the selection criteria were those markers with a p-value less than $1x10^{-3}$, leading to a set of 461 SNPs.Three experiments were conducted including the new set of variants: (1) Model of misclassified samples, (2) Mixed model with specific pre-selection by group and (3) Mixed model with variables resulting from previous models. All the experiments were carried out under the same scheme using cross-validation with 80% of the dataset for training and 20% for testing, and 10-fold CV for the lambda hyperparameter tuning.



(a) FS using solutions generated with Top100-CCun dataset. From 14 models generated, the best was model 13 with a maximum accuracy of 0.72 (black thick line.)

(b) FS using solutions generated with Top100-CCun dataset. From 13 models generated, the best was model 5 with a maximum accuracy of 0.7264 (black thick line.)

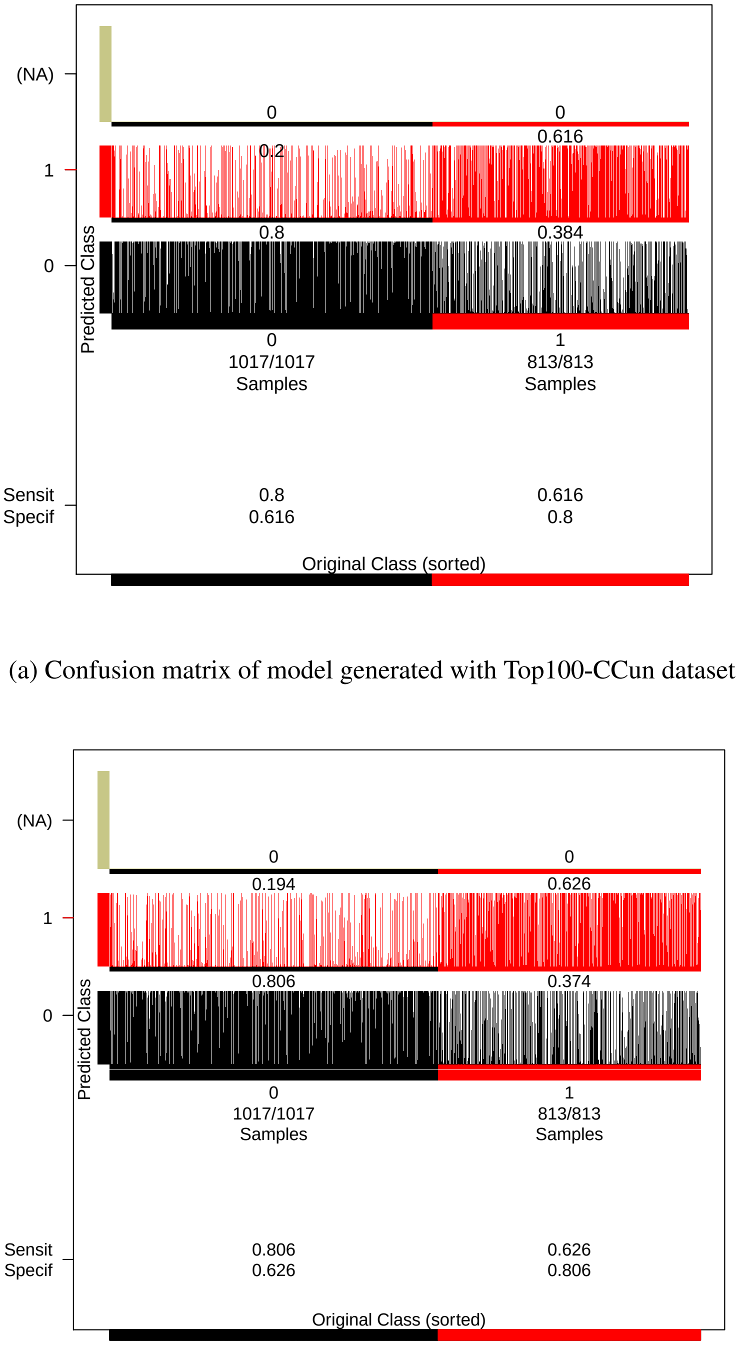
Figure 6.5: Forward selection using the most frequent markers. Solid line represents the overall accuracy and dashed lines represent the accuracy per class.

Model of misclassified samples

For this experiment, the 89 misclassified samples and the 461 pre-selected SNPs were taken to build a LASSO model. The results showed an accuracy of 0.9286, a sensitivity of 0.8571, a specificity of 0.9524 and an AUC of 0.9048. These results confirm that there are specific polymorphisms in this subgroup of individuals. Figure 6.8 shows the ROC curve of the model. The results show that there are specific markers in the subgroup of poorly classified samples that are not considered in the first run of the univariate analysis. However, when performing a second pre-selection in those misclassified examples, the performance of the model can be increased because now it includes relevant information.

Mixed model with specific pre-selection by group

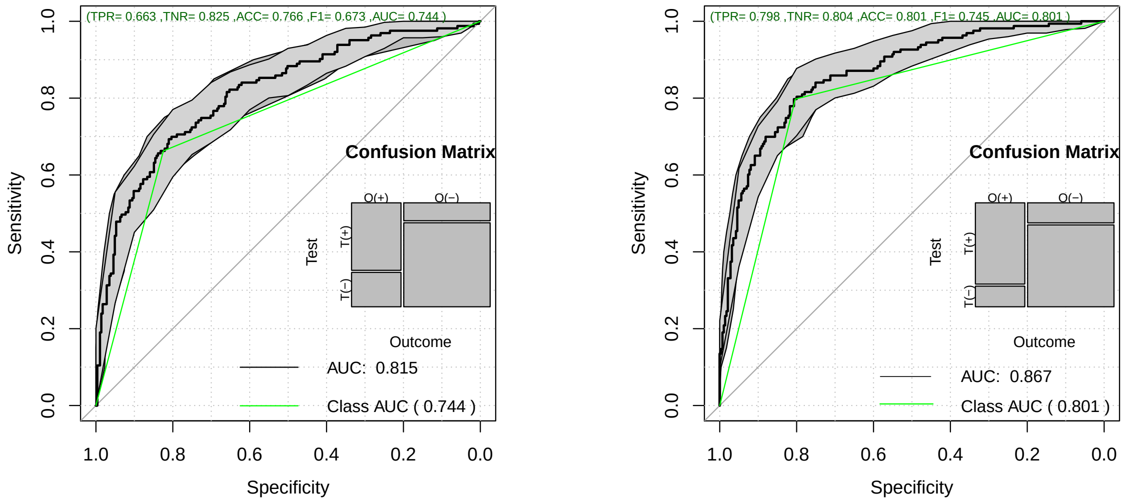
This experiment consisted of adding the 461 variants to the pre-selected 1,106 of the original model and built a new model with 1,567 input features. The overall performance of the new model did not showed improvements achieving an accuracy of 0.797, a sensitivity of 0.816, a specificity of 0.786 and an AUC of 0.801. Figure 6.9 shows the performance with the ROC curve of the model.



(a) Confusion matrix of model generated with Top100-CCun dataset.

(b) Confusion matrix of model generated with Th-CCun dataset.

Figure 6.6: Overall classification accuracy. Red: cases, black: controls.



(a) LASSO model with Top100-CCun dataset.

(b) LASSO model with Th-CCun dataset.

Figure 6.7: ROC curves and confusion matrices for LASSO models. Black line: continuous prediction, green line: operation point.

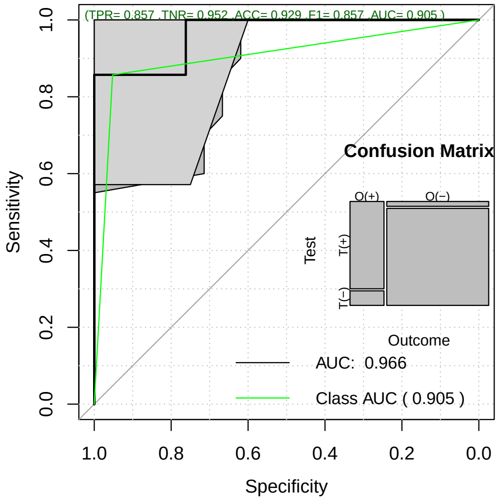


Figure 6.8: ROC curves and confusion matrices of model of misclassified samples. Black line: continuous prediction, green line: operation point.

Mixed model with variables resulting from previous models

The last experiment consisted of using 482 variables of the previous models (433 of the model Th-CCun and 49 of the model of poorly classified samples). The generated model of 358 variants had a performance that surpassed all previous models with an accuracy of 0.8438, a sensitivity of 0.8344, a specificity of 0.8491 and an AUC of 0.8417. Figure 6.10 shows the performance with the ROC curve of the model. The results show that there is a bias in the pre-selection with p-values, the first run of the univariate analysis does not manage to

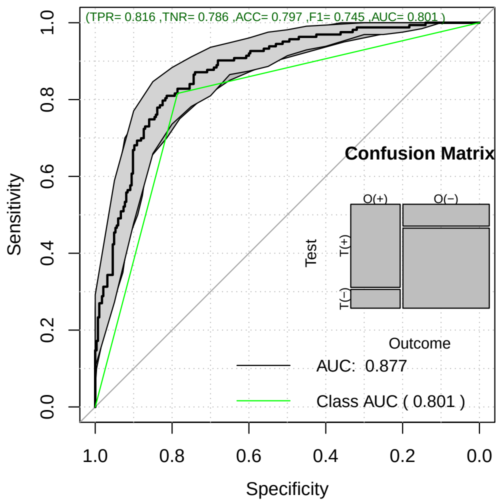


Figure 6.9: ROC curves and confusion matrices of model of mixed model with specific preselection by group. Black line: continuous prediction, green line: operation point.

capture the information of the small subgroup, which leads to a lower classification. However, by performing a second evaluation and including the variants resulting from the previous models, a new model with higher predictive power is obtained since important information of the population that was not included in the first model now is incorporated.

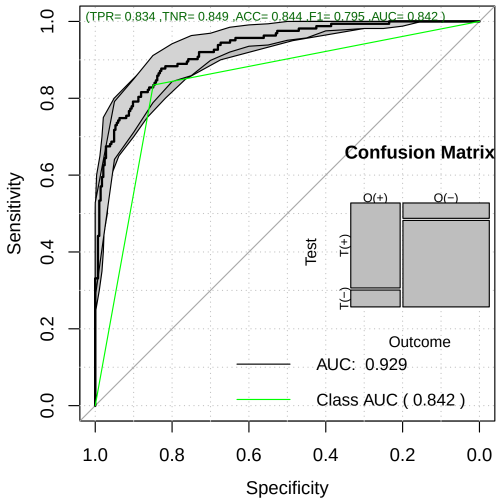


Figure 6.10: ROC curves and confusion matrices of model of mixed model with variables resulting from previous models. Black line: continuous prediction, green line: operation point.

Model	Dataset	Input features	Model length	ACC	Sensitivity	Specificity	AUC
BSWIMS	Top100-Ccun	100	32	0.6863	0.6262	0.7337	0.68
	Th-Ccun	1,106	55	0.699	0.6085	0.7704	0.6895
GA+SVM+FS	Top100-Ccun	100	49	0.72	0.616	0.8	0.708
	Th-Ccun	1,106	42	0.7264	0.626	0.806	0.716
LASSO	Top100-Ccun	100	71	0.7656	0.6626	0.8246	0.7436
	Th-Ccun	1,106	433	0.8013	0.7975	0.8035	0.8005
	Th-Misclassified	461	49	0.9286	0.8571	0.9524	0.9048
	Mix from pre-selection	1,567	493	0.7969	0.816	0.786	0.801
	Mix from previous models	**482**	**358**	**0.8438**	**0.8344**	**0.8491**	**0.8417**

Table 6.3: Summary of the characteristics and performance of the models.

6.3 Summary

This chapter summarized the results of the experiments carried out with the wrapper and embedded methods. The wrapper techniques included the BSWiMS algorithm and an ensemble of GA+SVM+FS while the embedded approach used was LASSO. The models were created with the two datasets Top100-CCun and Th-CCun so in total six experiments were assessed with confusion matrices, ROC curves, and heatmaps. LASSO with the mixture of variants from previous models was the model that outperformed the others with an AUC of 0.9124 and showed the best discrimination between classes.

7 | Model Analysis and Biological Relevance

In this chapter, the performance of the classification is evaluated by Principal Component Analysis, which represents the relationship of the samples using the SNPs selected from each representative model. Moreover, the identification of the genes associated with each variant of the models is presented. The search of the markers was done through the web page dbSNP of the NCBI [39]. Figure 7.1 shows an example of the genomic region of SNP rs405509 on chromosome 19 belonging to the APOE gene that has been previously associated with Alzheimer's disease.

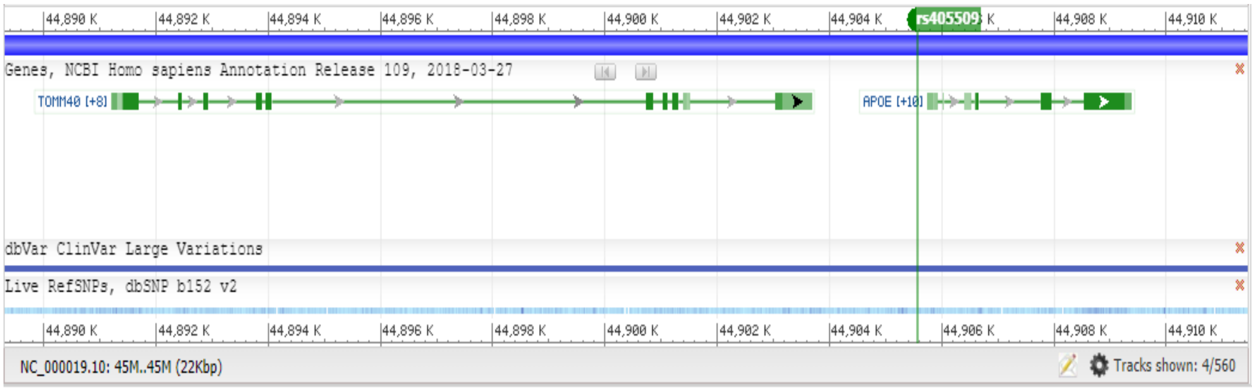

44,890 K
44,892 K
44,894 K
44,896 K
44,898 K
44,900 K
44,902 K
44,904 K
rs405509
44,906 K
44,908 K
44,910 K
Genes, NCBI Homo sapiens Annotation Release 109, 2018-03-27
TOMM40 [+8]
APOE [+10]
dbVar ClinVar Large Variations
Live RefSNPs, dbSNP b152 v2
NC_000019.10: 45M..45M (22Kbp)
Tracks shown: 4/560

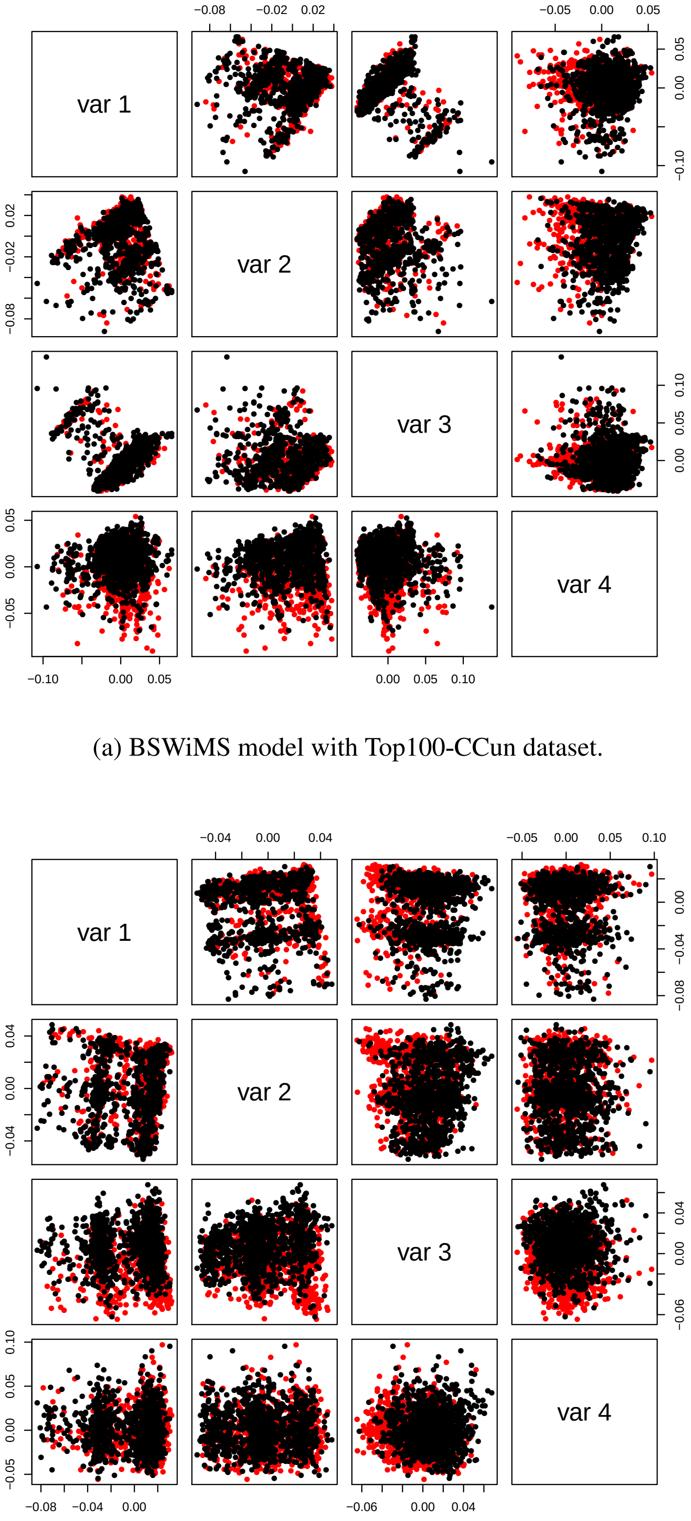
Figure 7.1: Genomic region of variant rs405509 on chromosome 19 belonging to the APOE gene.

7.1 BSWiMS models

A representation of the models in the PCA space is shown in figure 7.2. The discrimination is poor, only PC3 and PC4 of both models show a slight division between cases and controls given the selected markers in each one. Tables 7.2 and 7.2 list the markers of each model with their respective genes. As expected, there is a strong influence of chromosome 19 markers that have been previously associated with AD in other studies.

Name	Chr	Position	Gene
rs10203407	2	66471664	MEIS1
rs11124097	2	106262145	UXS1
rs7600624	2	152425861	FMNL2
rs7591606	2	152517361	FMNL2
rs4103004	3	20001342	PP2D1
rs230489	4	102467284	NFKB1
rs12215131	6	47741960	ADGRF4
rs3734444	6	55874755	BMP5
rs10104063	8	15707617	TUSC3
rs918245	8	15738578	TUSC3
rs11203752	8	16043710	MSR1
rs1921716	8	43402505	POTEA
rs4131198	8	43808532	POTEA
rs8175350	8	43894583	LOC105379397
rs1975804	8	108279004	LOC105375704
rs1789964	8	108514499	EMC2
rs1679666	8	108527618	EMC2
rs7070745	10	3163553	PITRM1
rs11248442	10	123200098	LOC107984275
rs1368008	14	58708697	DACT1
rs6574721	14	82839852	LINC02301
rs1543422	14	98716434	C14orf177
rs8062743	16	17285660	XYLT1
rs11861415	16	17290667	XYLT1
rs17683483	17	62725122	MARCH10
rs440277	19	44857967	NECTIN2
rs6859	19	44878777	NECTIN2
rs157580	19	44892009	TOMM40
rs2075650	19	44892362	TOMM40
rs8106922	19	44898409	TOMM40
rs405509	19	44905579	APOE
rs439401	19	44911194	APOC1

Table 7.1: Summary of genomic data for the BSWiMS model generated with Top-CCun dataset.



(a) BSWiMS model with Top100-CCun dataset.

(b) BSWiMS model with Th-CCun dataset.

Figure 7.2: Depiction of the BSWiMs models in PCA space. Red dots: cases, black dots: controls.

SNP	Chr	Position	Gene
rs11249395	1	121570001	EMBP1
rs2419117	1	168619944	LOC105371604
rs385341	2	40173349	SLC8A1
rs2110743	2	40311828	SLC8A1
rs741477	2	64839177	LINC01800
rs10203407	2	66471664	MEIS1
rs17213510	2	106126322	UXS1
rs7591606	2	152517361	FMNL2
rs2676510	2	172854792	RAPGEF4
rs3821236	2	191038032	STAT4
rs13383860	2	212389206	ERBB4
rs4103004	3	20001342	PP2D1
rs4677353	3	73849363	LINC02005
rs195098	4	5423014	STK32B
rs2702414	4	178478369	LOC105377563
rs1395095	5	10132851	FAM173B
rs10513044	5	10149113	FAM173B
rs1025003	5	170217841	C5orf58
rs2274154	6	33889748	LINC01016
rs12215131	6	47741960	ADGRF4
rs3734444	6	55874755	BMP5
rs9478967	6	151167181	AKAP12
rs7806237	7	147222997	CNTNAP2
rs330911	8	9138763	PPP1R3B
rs10104063	8	15707617	TUSC3
rs11203752	8	16043710	MSR1
rs1975804	8	108279004	LOC105375704
rs1789964	8	108514499	EMC2
rs1679666	8	108527618	EMC2
rs10964392	9	20025249	SLC24A2
rs10967072	9	25775264	LINC01241
rs1147808	9	25775449	LINC01241
rs10818288	9	119226621	BRINP1
rs7070745	10	3163553	PITRM1
rs1890583	10	3548077	LOC105376360
rs4881291	10	4108859	LOC107984196
rs4275521	10	4177141	LINC00702
rs7925016	11	62276457	SCGB2A2
rs7995077	13	23253676	SGCG
rs4142312	13	43000850	LOC105370179
rs1543422	14	98716434	C14orf177
rs2045498	16	17278824	XYLT1
rs11861415	16	17290667	XYLT1

Table 7.2 continued from previous page

SNP	Chr	Position	Gene
rs10871304	16	75189593	CTRB2
rs4783035	16	82455879	LOC101928392
rs17683483	17	62725122	MARCH10
rs9951431	18	29788820	LOC105372045
rs440277	19	44857967	NECTIN2
rs6859	19	44878777	NECTIN2
rs157580	19	44892009	TOMM40
rs2075650	19	44892362	TOMM40
rs8106922	19	44898409	TOMM40
rs405509	19	44905579	APOE
rs6086075	20	7515918	MIR8062
rs6078976	20	13290118	TASP1

Table 7.2: Summary of genomic data for the BSWiMS model generated with Th-CCun dataset.

7.2 GA+SVM+FS models

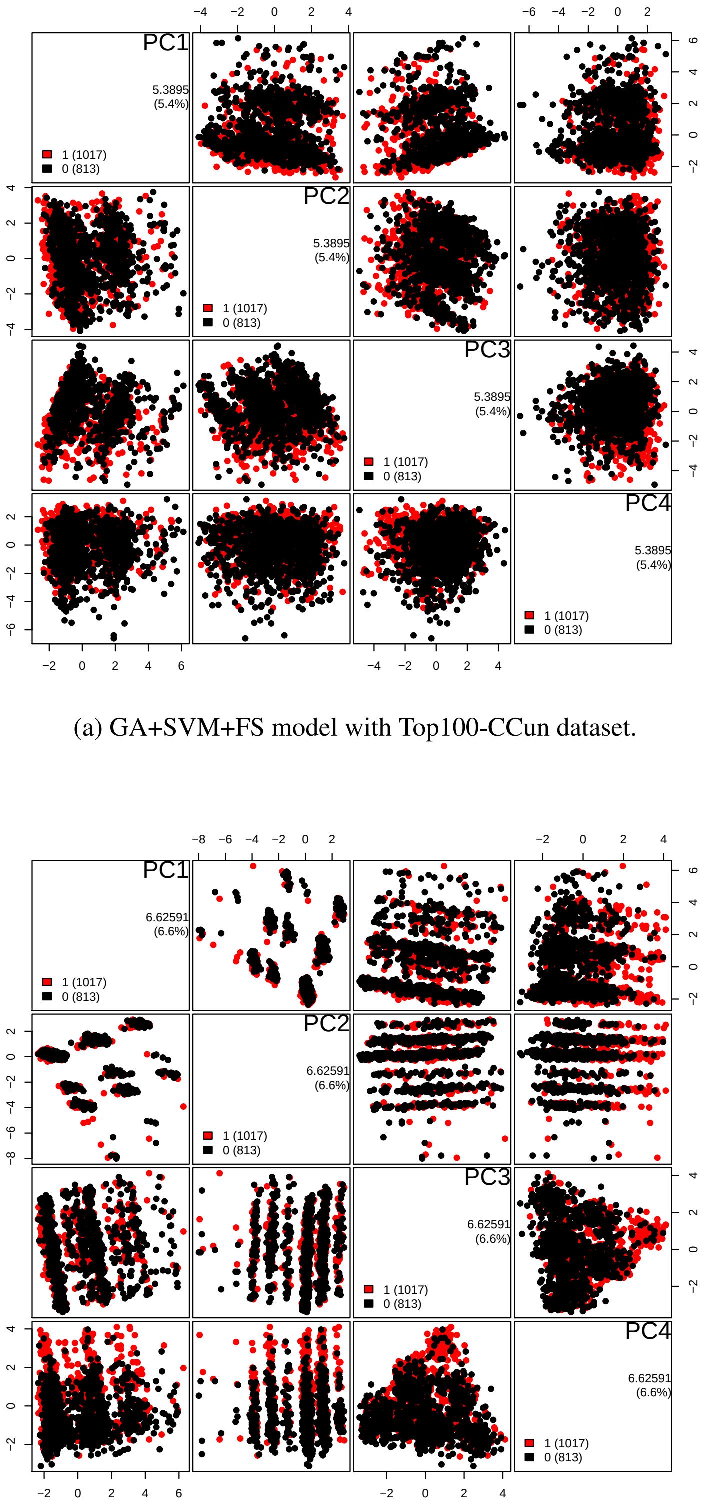
Figure 7.3 shows the performance of the models through a PCA of individuals. The model derived from Top100-CCun performs poorly only PC3 vs. PC4 show a slight separation of classes. Despite its similarity in terms of accuracy and AUC, the model generated with Th-CCun shows better discrimination with PC1 vs. PC4, PC2 vs. PC4, and PC3 vs. PC4.

7.3 LASSO models

Figure 7.4 illustrates a representation of the models in PCA space, where the model generated with Th-CCun makes better discrimination between cases and controls compared to the previous models.

7.3.1 Model of misclassified samples

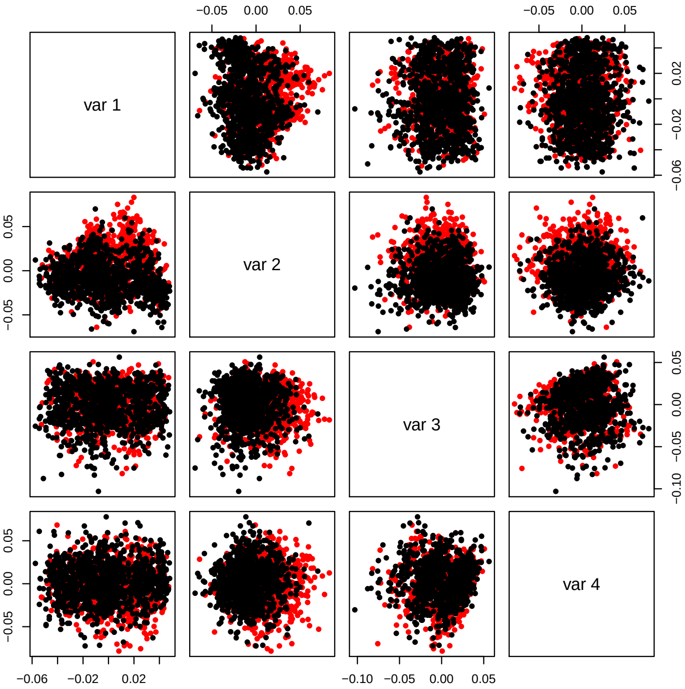
The space representation in the PCA of figure 7.5 shows that PC1 vs PC2, PC1 vs PC3, and PC1 vs PC4 make almost a complete separation between cases and controls. This result confirms that the 49 polymorphisms of the model can characterize the subset of poorly classified samples.



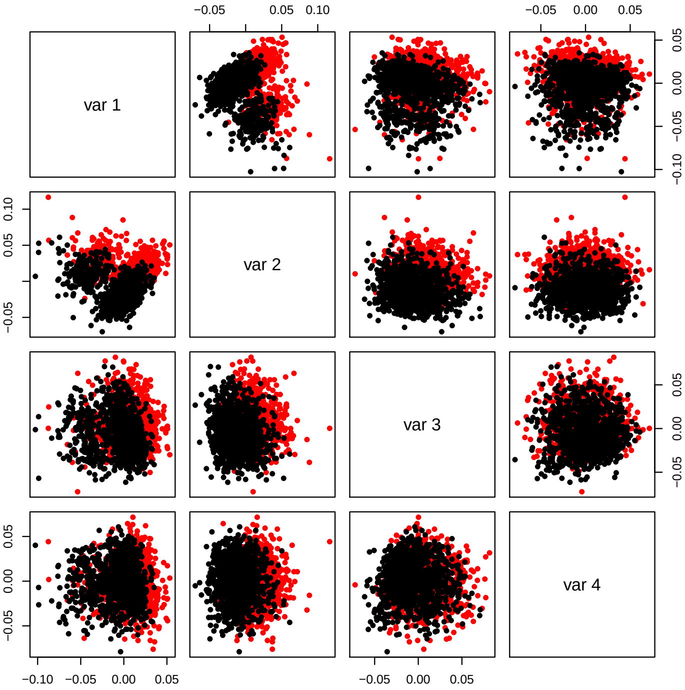
(a) GA+SVM+FS model with Top100-CCun dataset.

(b) GA+SVM+FS model with Th-CCun dataset.

Figure 7.3: Depiction of the GA+SVM+FS models in PCA space. Red dots: cases, black dots: controls.



(a) LASSO model with Top100-CCun dataset.



(b) LASSO model with Th-CCun dataset.

Figure 7.4: Depiction of the model in PCA space. Red dots: cases, black dots: controls.

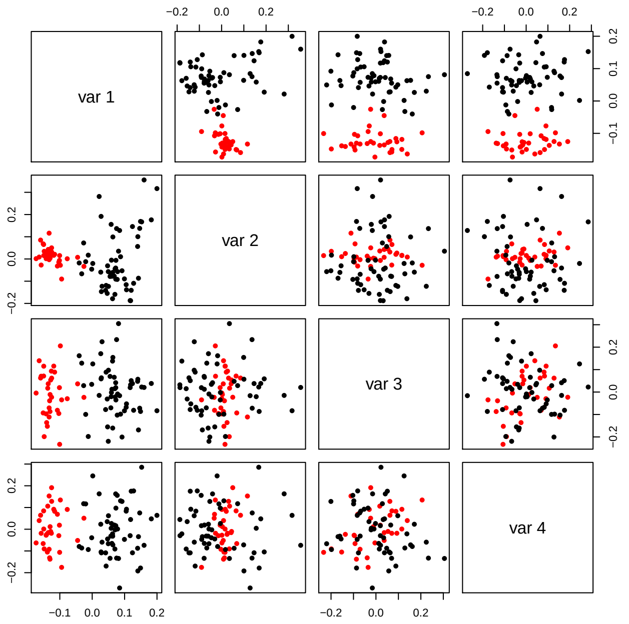


Figure 7.5: Depiction of the model of misclassified samples in PCA space. Red dots: cases, black dots: controls.

7.3.2 Mixed model with specific pre-selection by group

The results of this model showed discrimination between cases and controls quite similar to that made with the model of 1,106 SNPs, as can be seen in figure 7.6. One possible reason for this result is because the 1,567 variables that entered into the model competed with each other, and the model selected 493 markers that could have been previously chosen by the model of 433 variants, so there was no increase in performance.

7.3.3 Mixed model with variables resulting from previous models

Figure 7.7 shows a representation of the models in PCA space, where the mixed model generated with 482 variables from previous models. PC1 vs PC2, PC1 vs PC3, and PC1 vs PC4 are the components that show better discrimination between cases and controls compared to the previous models.

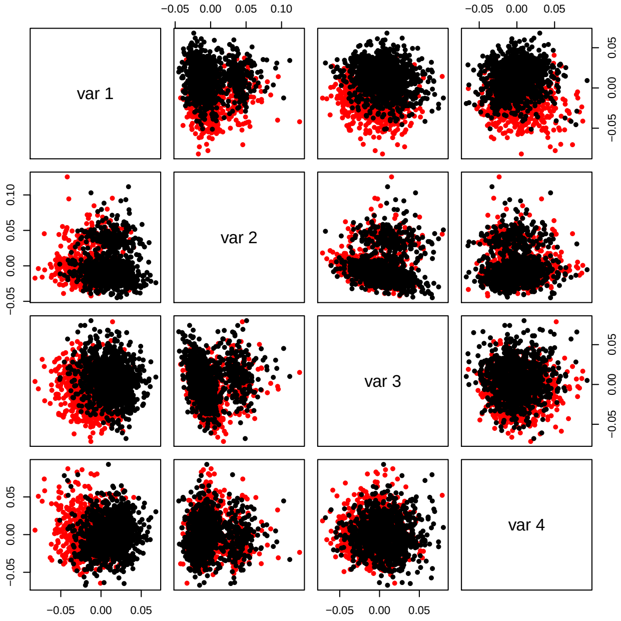


Figure 7.6: Depiction of the mixed model with specific pre-selection by group in PCA space. Red dots: cases, black dots: controls.

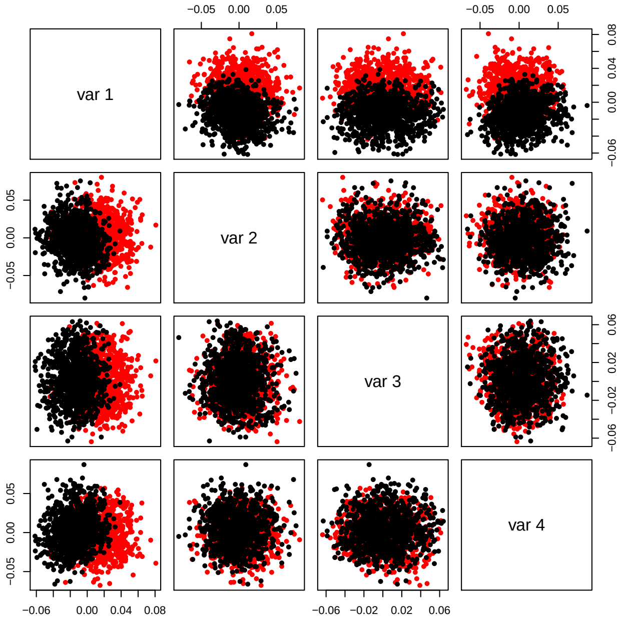


Figure 7.7: Depiction of the mixed model with variables resulting from previous models in PCA space. Red dots: cases, black dots: controls.

7.4 Summary

This chapter presented a way to analyze the models through a PCA of individuals. The models generated with LASSO achieved better discrimination between cases and controls. The genes of each of the variants of the models were also shown.

8 | Conclusions

This chapter presents the conclusions of the study, summarizing the experiments that were carried out and the results obtained. The genetic data went through quality control filters that removed variants that increased false positives and false negatives. The procedure was reviewed at each step with the results of the study conducted by Wijsman [17] as a guide since the data source is the same as in this research. The chi-squared test with kinship correction was also evaluated by a positive control to validate the computed p-values. The results showed a high correlation ($R^2 = 0.99$) compared with those of Wijsman. The pre-selection of features allowed the creation of two datasets. The first one included the best one hundred markers and 1,830 European-American individuals and was named Top100-CCun. The second one called Th-CCun included markers with a p-value lower than 1×10^-3 (1,106 SNPs) and the group of European-Americans.

The construction of the models included two strategies for the generation of combinations of markers, one by a step-wise method and another with genetic algorithms. BSWiMS was the step-wise method used, and it generated logistic models with similar results. The model built with the Top100-CCun dataset included 32 markers and achieved an AUC of 0.68. On the other hand, the second model based on Th-CCun had 55 SNPs and an AUC of 0.689. The representation of the models by PCA showed poor discrimination. The stochastic search was carried out with genetic algorithms coupled with SVM to obtain a set of solutions that subsequently generated a representative model through a forward selection procedure that added the SNPs by frequency. The markers of the Top100-CCun model could converge in the 50 solutions because the search space with 100 SNPs is smaller than the one with 1,106 predictors and due to the stochastic search the markers are repeated more frequently, unlike those of Th-CCun which needed 1,000 solutions to be stable. The performance of the models did not showed a big difference between both datasets, Top100-CCun and Th-CCun obtained AUCs of 0.708 and 0.716, respectively.

The embedded method LASSO was used to generate a predictive model. This algorithm has already been used in previous works of disease prediction because of its feature selection stage within the model training. The model of 433 markers generated with Th-CCun outperformed the others with an AUC of 0.8005. The Top100-CCun model with 71 features obtained an AUC of 0.7436. The results suggest that the AUC increases as the number of features increases. In the PCA representation, again the Th-CCun model showed the best discrimination between classes.

To the best of our knowledge, this is the first study that uses step-wise methods and genetic algorithms based on GWAS data to explore the search space efficiently identifying those variants with a higher association with Alzheimer's Disease. Also, unlike other researches present in the literature, this study incorporates the analysis of poorly classified samples in predictive models to increase prediction successfully.

Unlike most studies that only assess the performance of the models with data analytics metrics such as accuracy or AUC, this study also explains the biological interpretation by identifying the genes to which the variants belong. As expected, the models included markers previously associated with AD, mainly in the genomic region of chromosome 19 near APOE. A more detailed biological interpretation is suggested for future studies.

Overall, the models performed well despite being based solely on genetic data. As already mentioned, AD has a mixture of hereditary, environmental and lifestyle factors. Further experiments include adding biological information and clinical data such as magnetic resonances or cognitive studies, to increase the accuracy of the models. Another improvement suggested is to analyze more thoroughly the misclassified variants in the models. The use of multivariate methods allowed to consider the interaction effects between the markers, unlike univariate methods. Machine learning approaches successfully generated predictive models that include combinations of DNA polymorphisms associated with Alzheimer's disease.

9 | Appendixes

9.1 Appendix A

Tables 9.1 and 9.2 show in detail the quality control criteria and the markers that were removed for each chromosome.

											SNPs by Chromosome												
Chromosome	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	**TOTAL**
Initial markers	47,108	49,517	40,873	36,812	37,416	42,665	34,080	33702	28632	32079	29711	29330	22693	19930	18632	18816	16391	17817	11034	15095	8853	9284	600470
Markers without reference SNP ID number (“rs” ID)																							
Removed markers	1447	1364	946	1015	1212	1466	1315	887	644	1043	777	582	455	573	782	963	745	369	556	327	217	246	17931
Remaining markers	45661	48153	39927	35797	36204	41199	32765	32815	27988	31036	28934	28748	22238	19357	17850	17853	15646	17448	10478	14768	8636	9038	582539
Call Rate <98%																							
Removed markers	1091	1146	781	915	834	1242	1047	710	643	787	705	659	485	439	552	491	483	367	363	291	231	326	14588
Remaining markers	44570	47007	39146	34882	35370	39957	31718	32105	27345	30249	28229	28089	21753	18918	17298	17362	15163	17081	10115	14477	8405	8712	567951
Monomorphic markers																							
Removed markers	1043	975	764	686	676	1039	723	583	466	594	611	635	377	344	401	421	374	319	236	284	133	226	11910
Remaining markers	43527	46032	38382	34196	34694	38918	30995	31522	26879	29655	27618	27454	21376	18574	16897	16941	14789	16762	9879	14193	8272	8486	556041
Hardy Weinberg (p<1x10^-5)																							
Removed markers	60	84	74	56	49	118	44	44	57	32	48	38	41	29	33	28	34	27	27	24	16	14	977
High-quality markers	43467	45948	38308	34140	34645	38800	30951	31478	26822	29623	27570	27416	21335	18545	16864	16913	14755	16735	9852	14169	8256	8472	555064

Table 9.1: Summary of quality control procedures applied to control markers.

QC Criterion	SNPs by Chromosome																					
\textbf{Chromosome}	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	TOTAL
\textbf{Initial markers}	47108	49517	40873	36812	37416	42665	34080	33702	28632	32079	29711	29330	22693	19930	18632	18816	16391	17817	11034	15095	8853	9284	600470
Markers without reference SNP ID number (“rs” ID)																							
Removed markers	1447	1364	946	1015	1212	1466	1315	887	644	1043	777	582	455	573	782	963	745	369	556	327	217	246	17931
Remaining markers	45661	48153	39927	35797	36204	41199	32765	32815	27988	31036	28934	28748	22238	19357	17850	17853	15646	17448	10478	14768	8636	9038	582539
Call Rate <98%																							\textbf{}
Removed markers	1091	1146	781	915	834	1242	1047	710	643	787	705	659	485	439	552	491	483	367	363	291	231	326	14588
Remaining markers	44570	47007	39146	34882	35370	39957	31718	32105	27345	30249	28229	28089	21753	18918	17298	17362	15163	17081	10115	14477	8405	8712	567951
Monomorphic markers																							
Removed markers	1016	1094	735	849	781	1201	1002	686	627	750	682	600	467	411	532	476	454	361	357	270	208	298	13857
Remaining markers	44645	47059	39192	34948	35423	39998	31763	32129	27361	30286	28252	28148	21771	18946	17318	17377	15192	17087	10121	14498	8428	8740	568682
Hardy Weinberg (p\textless 1x10-5)																							
Removed markers	1028	896	670	622	521	948	683	535	423	506	534	583	409	274	364	330	365	284	220	273	121	207	10796
High-quality markers	43617	46163	38522	34326	34902	39050	31080	31594	26938	29780	27718	27565	21362	18672	16954	17047	14827	16803	9901	14225	8307	8533	557886

Table 9.2: Summary of quality control procedures applied to cases markers.

9.2 Appendix B

List of 71 markers of the LASSO model with the Top100-CCun dataset.

1. rs495959
2. rs10180600
3. rs10203407
4. rs10490103
5. rs11124097
6. rs1372140
7. rs1568538
8. rs2577180
9. rs3810831
10. rs3820852
11. rs6434081
12. rs6545286
13. rs6548077
14. rs7591606
15. rs938487
16. rs11707751
17. rs4103004
18. rs9867793
19. rs230489
20. rs11759748
21. rs12215131
22. rs226930
23. rs2328869
24. rs3734444
25. rs9345943
26. rs9354514
27. rs10278342
28. rs1294906
29. rs13241213
30. rs10104063
31. rs11203752
32. rs12544662
33. rs1679666
34. rs1975804
35. rs4131198
36. rs8175350
37. rs918245
38. rs11248456
39. rs7070745
40. rs10838333
41. rs12420859
42. rs2156469
43. rs4757064
44. rs2695148
45. rs7136901
46. rs3129599
47. rs1368008
48. rs1543422
49. rs17676330
50. rs6574721
51. rs17697225
52. rs723804
53. rs11861415
54. rs1466134
55. rs2404691
56. rs8062743
57. rs17683483
58. rs4796606
59. rs561811
60. rs7222487
61. rs768027
62. rs9965853
63. rs157580
64. rs2075650
65. rs405509
66. rs439401
67. rs440277
68. rs6859
69. rs714948
70. rs8103315
71. rs8106922

List of 433 markers of the LASSO model with the Th-CCun dataset. Character + means that the marker is present in the mixed model with variables resulting from previous models.

1. rs10493425 +
2. rs10494179 +
3. rs10494824
4. rs10923633 +
5. rs11208148
6. rs11240574 +
7. rs11577496 +
8. rs11580561
9. rs12046077 +
10. rs12070470 +
11. rs12094388 +
12. rs12132851
13. rs1321106 +
14. rs1674877 +
15. rs1693230 +
16. rs17375108 +
17. rs1778037 +
18. rs2025689 +
19. rs2027278 +
20. rs2066298 +
21. rs2367270 +
22. rs2419117 +
23. rs2816065 +
24. rs281968 +
25. rs3009577 +
26. rs3789541 +
27. rs4650571 +
28. rs4658104 +
29. rs4838937
30. rs6541080 +
31. rs6594018
32. rs6677245 +
33. rs6677436 +
34. rs6690148 +
35. rs7518469 +
36. rs7531902 +
37. rs7534822 +
38. rs922305 +
39. rs10180600
40. rs10203407 +
41. rs10490103 +
42. rs105287 +
43. rs10928631 +
44. rs11692473 +
45. rs11889094
46. rs1214118 +
47. rs12616481 +
48. rs13383860 +
49. rs1372140 +
50. rs1507670 +
51. rs1551728
52. rs1568538
53. rs16849846 +
54. rs17213510 +
55. rs192634 +
56. rs2086720
57. rs2110743 +
58. rs2139246
59. rs2160930 +
60. rs2163050
61. rs2577180 +
62. rs2676510 +
63. rs2676517
64. rs354230 +
65. rs3750272 +
66. rs3768651 +
67. rs3770560 +
68. rs3795859
69. rs3821236 +
70. rs4552231 +
71. rs4832054 +
72. rs6548077 +
73. rs7419384 +
74. rs751192 +
75. rs7559029 +
76. rs7581030 +
77. rs7601596
78. rs775800 +
79. rs938487
80. rs1152219
81. rs11718528
82. rs124844 +
83. rs13091847 +
84. rs1365152
85. rs1525011
86. rs16861384 +
87. rs17220361 +
88. rs17386464 +
89. rs17646659 +
90. rs2315966 +
91. rs4074707 +
92. rs4103004 +
93. rs4677353 +
94. rs4839614 +
95. rs657079 +
96. rs658395
97. rs6766673 +
98. rs6780036 +
99. rs6796391 +
100. rs6804202 +
101. rs6809414 +
102. rs7374825 +
103. rs7613141 +
104. rs7625222 +
105. rs9311717
106. rs9812253 +
107. rs9838117 +
108. rs10012946 +
109. rs10516213 +
110. rs11721741
111. rs11725488
112. rs12506292 +
113. rs1508986
114. rs1567396 +
115. rs16874378 +
116. rs17040984 +
117. rs1830310 +
118. rs195098 +
119. rs230489 +
120. rs2702414 +
121. rs2732204 +
122. rs3913195
123. rs6842379 +
124. rs746977 +
125. rs7667609
126. rs9312333 +
127. rs9991949 +
128. rs9997024
129. rs10038395 +
130. rs10062689
131. rs1025003 +
132. rs10463996
133. rs10513044 +
134. rs11743438
135. rs11746232 +
136. rs11748934 +
137. rs11750465 +
138. rs12173156 +
139. rs16895144 +
140. rs17064682

141. rs17555606 +
142. rs1833755 +
143. rs244436 +
144. rs2471042
145. rs25952
146. rs27911
147. rs40474
148. rs4382202 +
149. rs4425488
150. rs4835811
151. rs6862189 +
152. rs6870664 +
153. rs7708940
154. rs7731517
155. rs918507 +
156. rs9312858
157. rs9312938 +
158. rs10948259 +
159. rs11155364 +
160. rs12215131 +
161. rs1890061 +
162. rs2024694 +
163. rs2130493
164. rs226930 +
165. rs2274154 +
166. rs2397060 +
167. rs269998
168. rs2764092 +
169. rs2881062
170. rs3734444 +
171. rs3761986 +
172. rs4288220
173. rs4712455 +
174. rs614002
175. rs7748472
176. rs7773694
177. rs9275615
178. rs9276162 +
179. rs9347691 +
180. rs9356457 +
181. rs9375645
182. rs9385647 +
183. rs9386053 +
184. rs9404549 +
185. rs9445243 +
186. rs9460121
187. rs9460196 +
188. rs9475436 +
189. rs9478967 +
190. rs9493022 +
191. rs10276579 +
192. rs1063964 +
193. rs12154478 +
194. rs12539196 +
195. rs13240249
196. rs1806453 +
197. rs219236 +
198. rs3735108 +
199. rs3735109 +
200. rs4722614 +
201. rs4727332 +
202. rs6961482 +
203. rs715851 +
204. rs723340 +
205. rs7781339 +
206. rs7782538
207. rs7795485 +
208. rs7806237 +
209. rs10104063 +
210. rs10104107 +
211. rs12544429 +
212. rs12544662 +
213. rs12549180 +
214. rs12682103 +
215. rs13267988 +
216. rs16906883 +
217. rs16937025 +
218. rs17303296 +
219. rs1972436 +
220. rs1975804 +
221. rs330911 +
222. rs352768 +
223. rs6474486 +
224. rs6991681 +
225. rs7008113 +
226. rs7015856 +
227. rs7386016 +
228. rs7832008 +
229. rs9644361 +
230. rs10125854 +
231. rs10815710 +
232. rs10818288 +
233. rs10967072 +
234. rs10981116 +
235. rs11795111
236. rs13288998
237. rs17520192 +
238. rs3847320 +
239. rs4242670 +
240. rs4255200
241. rs716738 +
242. rs10764786 +
243. rs10824233 +
244. rs10870341
245. rs10904757
246. rs11010758 +
247. rs11016073 +
248. rs11248442 +
249. rs11593607
250. rs11816291 +
251. rs12219789
252. rs12357398 +
253. rs12572759 +
254. rs1456285 +
255. rs1890583 +
256. rs2247977
257. rs2282367 +
258. rs2282372 +
259. rs2342606 +
260. rs2483578 +
261. rs2780685
262. rs2797142 +
263. rs4275521 +
264. rs4880336 +
265. rs4881291 +
266. rs5003944 +
267. rs615019
268. rs62209 +
269. rs6479815 +
270. rs7070745 +
271. rs7070797
272. rs7350421
273. rs756208 +
274. rs7898954 +
275. rs7922515 +
276. rs10743152 +
277. rs10895458 +
278. rs10896138 +
279. rs11033021
280. rs12273601 +
281. rs12420859 +
282. rs16932405
283. rs1783180 +
284. rs2291753 +
285. rs2939753
286. rs35852551
287. rs3847580
288. rs4261298 +
289. rs474951
290. rs4757064 +
291. rs523223 +
292. rs528823 +

293. rs6578752 +
294. rs7119775 +
295. rs7122505 +
296. rs7128017 +
297. rs7924909 +
298. rs7925016 +
299. rs7931399 +
300. rs7934021 +
301. rs11176056
302. rs11832405
303. rs1436849 +
304. rs1520782
305. rs16933825 +
306. rs17023308 +
307. rs17726837
308. rs17769793
309. rs201363 +
310. rs2287476 +
311. rs7136901 +
312. rs11619301
313. rs11841466 +
314. rs17070878 +
315. rs17288723 +
316. rs1924960 +
317. rs4142312 +
318. rs4255684 +
319. rs7335730 +
320. rs7995077 +
321. rs931810
322. rs9533767 +
323. rs9546945
324. rs10138354 +
325. rs10483438 +
326. rs11160024 +
327. rs11622299 +
328. rs1241513 +
329. rs1543422 +
330. rs17676330
331. rs2145472 +
332. rs234577
333. rs6574721 +
334. rs743249 +
335. rs8010460
336. rs8021478 +
337. rs900360 +
338. rs10152725 +
339. rs1224657 +
340. rs12439636 +
341. rs12440317 +
342. rs12915921 +
343. rs1902131 +
344. rs2161936 +
345. rs2959823 +
346. rs371125 +
347. rs3803523 +
348. rs4293320 +
349. rs6494798
350. rs8030144
351. rs8035375 +
352. rs8041377 +
353. rs9806693 +
354. rs10445014 +
355. rs10871304 +
356. rs11644916
357. rs11647105 +
358. rs11861415 +
359. rs12708762
360. rs12719801 +
361. rs1466134 +
362. rs1482246 +
363. rs16945930 +
364. rs2404691
365. rs2744148
366. rs3859167 +
367. rs4238574 +
368. rs4780464
369. rs4780947 +
370. rs4783035 +
371. rs7188223 +
372. rs7192876
373. rs9746438 +
374. rs12150632 +
375. rs1530361 +
376. rs17683483 +
377. rs4789374 +
378. rs4796606 +
379. rs7207432 +
380. rs7210118
381. rs7215706 +
382. rs740559 +
383. rs8082635 +
384. rs10514087 +
385. rs1196581 +
386. rs12454547
387. rs12456174 +
388. rs12456720 +
389. rs1436907 +
390. rs1590106
391. rs17068252
392. rs17089368 +
393. rs1942196
394. rs4800995 +
395. rs4890490
396. rs4940643 +
397. rs635538 +
398. rs7240797
399. rs755776 +
400. rs8094970 +
401. rs9945969 +
402. rs9947925 +
403. rs9951236
404. rs9951431 +
405. rs9964228 +
406. rs9965853
407. rs157580 +
408. rs2075650 +
409. rs346763 +
410. rs3764551
411. rs439401 +
412. rs440277 +
413. rs6859 +
414. rs8106922 +
415. rs156617 +
416. rs159764 +
417. rs2762926 +
418. rs3843781 +
419. rs433279 +
420. rs6014216 +
421. rs6059033
422. rs6078976 +
423. rs6086075 +
424. rs6130940 +
425. rs16987537 +
426. rs2164171
427. rs2822740 +
428. rs372992 +
429. rs132323 +
430. rs17745316 +
431. rs3788528
432. rs5761163 +
433. rs9610841 +

List of 461 markers from the pre-selection by univariate analysis using misclassified samples. Character * means that the marker is present in the model of misclassified samples and character + means that the marker is present in the mixed model with variables resulting from previous models.

1. rs1033806
2. rs10492963
3. rs10492966
4. rs10802454
5. rs10913149
6. rs10913157
7. rs10918196
8. rs1110266
9. rs11121203
10. rs11121213
11. rs11121229
12. rs11121230
13. rs11805932
14. rs12137865
15. rs12410886
16. rs12410893
17. rs1256348
18. rs12567592
19. rs1353130
20. rs16853647
21. rs17663802
22. rs1876304
23. rs2146195
24. rs2364815
25. rs2502436
26. rs2896987
27. rs3766735 *
28. rs4656224
29. rs4838917
30. rs4908777
31. rs622721
32. rs6656249
33. rs6670505
34. rs6671659
35. rs6698385 *
36. rs7535432
37. rs7544572
38. rs7552832
39. rs953535
40. rs954276
41. rs10165187
42. rs10174307
43. rs10196277
44. rs10204426
45. rs10490227
46. rs11688315
47. rs11891417 *
48. rs11892543
49. rs12615485
50. rs13028070
51. rs13389636
52. rs1344105
53. rs1519896
54. rs16839653
55. rs16987893
56. rs17509739
57. rs1840139
58. rs1915170
59. rs2077790
60. rs2084713
61. rs2103127
62. rs2245633
63. rs2600791
64. rs2724929 * +
65. rs298916
66. rs3732191
67. rs4303750
68. rs4493303
69. rs4663003 *
70. rs6724315
71. rs6732237
72. rs7588167 * +
73. rs7590358
74. rs7591591
75. rs7601145
76. rs10460880
77. rs10936526
78. rs11719400
79. rs11720467
80. rs1520599
81. rs2421748
82. rs4955666
83. rs4955865
84. rs6549877 * +
85. rs7618166
86. rs7627506
87. rs9842344
88. rs9873710
89. rs10000266 * +
90. rs10003762
91. rs10008779
92. rs11727494
93. rs11940954
94. rs11942069
95. rs12233744
96. rs12640501
97. rs12649449
98. rs13117931
99. rs13121519
100. rs13128505
101. rs13134269 *
102. rs1435474
103. rs1471263
104. rs1477538
105. rs1566481 *
106. rs1606110
107. rs1874238
108. rs1881523
109. rs2169736
110. rs2198743
111. rs2319261
112. rs2319264
113. rs4001038
114. rs4269167
115. rs4283687
116. rs6856230
117. rs734312
118. rs7654693
119. rs7665794
120. rs7669089
121. rs7688628
122. rs7697495
123. rs941130
124. rs10079889
125. rs10900862
126. rs10900864
127. rs11750513
128. rs12188982
129. rs12652257
130. rs12654269
131. rs12656498
132. rs12659754
133. rs12691271
134. rs12716144
135. rs13156236
136. rs13436221

137. rs1348433
138. rs152048
139. rs156055
140. rs1564800
141. rs1678934
142. rs16880442
143. rs17056764
144. rs17135605
145. rs17303482
146. rs17358298
147. rs17490659
148. rs17824668
149. rs1867324
150. rs1916998 *
151. rs1946649
152. rs2008384
153. rs2091803
154. rs218041 * +
155. rs218042 *
156. rs2195365
157. rs2242223
158. rs2245869
159. rs2290987
160. rs2416315
161. rs255041
162. rs26990
163. rs2964589
164. rs36724
165. rs3924257
166. rs4333317
167. rs459743
168. rs469844
169. rs4835726
170. rs509237
171. rs6580152
172. rs680503
173. rs6896919
174. rs7719681
175. rs7722673
176. rs7727656
177. rs7733404
178. rs935162
179. rs969400
180. rs970456 * +
181. rs10484663
182. rs11968721
183. rs12214904
184. rs13209334
185. rs1452531
186. rs1452537
187. rs1528505
188. rs162279
189. rs17387507
190. rs2029965 *
191. rs236246
192. rs2493237
193. rs2636623 * +
194. rs2677829 * +
195. rs2816377 * +
196. rs2817090
197. rs2817114
198. rs2987583
199. rs330109
200. rs4895979
201. rs600188 * +
202. rs6455807
203. rs649612
204. rs683423
205. rs6922131
206. rs6936473 *
207. rs7356837
208. rs7741391 * +
209. rs7760369 *
210. rs7760647 *
211. rs807517
212. rs818276
213. rs828568
214. rs855379
215. rs9320813 *
216. rs9382183
217. rs9397340
218. rs9477166 * +
219. rs9479642
220. rs9479658
221. rs9490199
222. rs10228095
223. rs10273276
224. rs1528500
225. rs1557968
226. rs17146301
227. rs17153151
228. rs1913409
229. rs2392376
230. rs2709778 * +
231. rs2867161
232. rs2894523
233. rs3912067
234. rs4725537
235. rs4727720
236. rs764513
237. rs7788516
238. rs10956454
239. rs12545387
240. rs13251215
241. rs16920503
242. rs17348854
243. rs2028223
244. rs2280849
245. rs4379435
246. rs4871799
247. rs6651216
248. rs6984752
249. rs6997005
250. rs7008223
251. rs7010943
252. rs892353
253. rs10521449
254. rs10758127
255. rs10758988
256. rs10818708 * +
257. rs10970899
258. rs11139983 *
259. rs1329383
260. rs2150637
261. rs2777793
262. rs3095748
263. rs663838
264. rs676113 * +
265. rs7866103
266. rs7873885
267. rs1033962
268. rs1033963
269. rs10437447
270. rs10827266
271. rs10828594 * +
272. rs10886772
273. rs10997144
274. rs11199574 * +
275. rs16916984
276. rs1911318
277. rs2050558
278. rs2131265
279. rs2256432
280. rs2797577
281. rs2804479 *
282. rs4319409
283. rs518525
284. rs649611
285. rs7070107
286. rs7075768
287. rs7089879
288. rs787667

289. rs7903736
290. rs947260
291. rs11225633
292. rs11602155
293. rs12223678
294. rs12417249
295. rs12577323
296. rs1372269
297. rs16928442
298. rs17374415
299. rs17686304 *
300. rs1792218
301. rs1811815
302. rs1941399
303. rs2445192
304. rs2509094
305. rs2852425
306. rs2852447
307. rs4550176
308. rs577625
309. rs6592522
310. rs6592554
311. rs7113370
312. rs7116059
313. rs7116191
314. rs7130671
315. rs739677
316. rs7934287
317. rs800336
318. rs947781
319. rs10748176 * +
320. rs10859071
321. rs11614654
322. rs12299204
323. rs12824958
324. rs388626 * +
325. rs4272843
326. rs4768967
327. rs597340 *
328. rs6580784
329. rs6580792 * +
330. rs1028965
331. rs1107169
332. rs11843993
333. rs149874
334. rs17552089
335. rs17792133
336. rs188014
337. rs200220
338. rs201763
339. rs201768
340. rs201803
341. rs2152580
342. rs4053534
343. rs796027
344. rs7985791
345. rs806312
346. rs881084
347. rs9316728
348. rs9526671
349. rs9539155 * +
350. rs9543383
351. rs9548957
352. rs1005564
353. rs10137082
354. rs10150022
355. rs1151573
356. rs11624780
357. rs12886510
358. rs12892905
359. rs17099852
360. rs1950968
361. rs1951294
362. rs1954333
363. rs2110706
364. rs2215134
365. rs2295174
366. rs3811178
367. rs4572291
368. rs4901761 * +
369. rs4905930
370. rs6575061
371. rs7142206
372. rs7151233
373. rs8003262
374. rs8013123
375. rs8020630
376. rs8022758
377. rs9285571 * +
378. rs12906076
379. rs1435683
380. rs1442291
381. rs17114595
382. rs2730085
383. rs335489 * +
384. rs335507
385. rs4775398
386. rs6598406
387. rs7170140
388. rs7497343
389. rs11117255 * +
390. rs12934725
391. rs17769799
392. rs3849233
393. rs3860272
394. rs4297685
395. rs8046199
396. rs9454 * +
397. rs11656744
398. rs12941303
399. rs12951391
400. rs16949616
401. rs16959714
402. rs16959820
403. rs16977009 * +
404. rs17767678
405. rs2058010
406. rs2158917
407. rs2529379
408. rs2871647
409. rs427488
410. rs4646342
411. rs4646344
412. rs4790904
413. rs4791474
414. rs740570
415. rs9907959
416. rs11659982
417. rs2542178
418. rs299739
419. rs299740
420. rs4986203
421. rs644601
422. rs8085108
423. rs10408331
424. rs1126757
425. rs11667291
426. rs12462703
427. rs12611090
428. rs1654654
429. rs2283575
430. rs2305744
431. rs2451996
432. rs35193259
433. rs380731
434. rs759048
435. rs8189791
436. rs8189845
437. rs8189858
438. rs860386
439. rs870379
440. rs1487328

441. rs1883147 *
442. rs1883521 *
443. rs2208464
444. rs292858
445. rs432448
446. rs535315
447. rs6016262 *
448. rs6025863
449. rs6028885
450. rs980984 * +
451. rs1032002
452. rs1475903
453. rs2051397
454. rs2070519
455. rs2250930 * +
456. rs2839509
457. rs45430
458. rs9325634
459. rs6008046
460. rs738865
461. rs738881

Bibliography

[1] LAURA AURIA, AND ROUSLAN A. MORO. Support Vector Machines (SVM) as a Technique for Solvency Analysis . *SSRN Electronic Journal 1* (2008).

[2] NATIONAL CENTRALIZED REPOSITORY FOR ALZHEIMER'S DISEASE AND RELATED DEMENTIAS. The Genetics of Alzheimer's Disease , 2015. `https://ncrad.iu.edu/genetics_ad.html`.

[3] WORLD HEALTH ORGANIZATION. Dementia, 2017. `https://www.who.int/news-room/fact-sheets/detail/dementia`.

[4] A. ALZUBAIDI AND G. COSMA. A multivariate feature selection framework for high dimensional biomedical data classification. In *2017 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)* (Aug 2017), pp. 1–8.

[5] A MARTINEZ-TORTEYA AND I ALANIS AND J TAMEZ-PENA. FeatuRE Selection Algorithms for Computer-Aided Diagnosis: an R package. *Submitted* (2018), .

[6] AHMET TURAN ISIK. Late onset Alzheimer's disease in older people. *Clinical Interventions in Aging 5*, 1 (2010), 307–311.

[7] ALKES L PRICE1, NICK J PATTERSON, ROBERT M PLENGE, MICHAEL E WEINBLATT, NANCY A SHADICK AND, DAVID REICH. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics 38*, 8 (2006), 904–909.

[8] ALPAYDIN, E. *Machine learning : the new AI.* MIT Press essential knowledge series. MIT Press, Cambridge, MA, 2016.

[9] BING HAN, XUE-WEN CHEN , ZOHREH TALEBIZADEH, HUA XU. Genetic studies of complex human diseases: Characterizing SNP-disease associations using Bayesian networks. *BMC Systems Biology 6*, 3 (2012).

[10] CARL A. ANDERSON. *Analysis of Complex Disease Association Studies.* Academic Press, USA, 2011.

[11] CHEN, HONGGE. *Novel machine learning approaches for modeling variations in semiconductor manufacturing*. PhD thesis, 01 2017.

[12] CHRISTINE HEROLD , MICHAEL STEFFENS, FELIX F. BROCKSCHMID , MAX P. BAUR AND TIM BECKER. INTERSNP: genome-wide interaction analysis guided by a priori information. *Bioinformatics 25*, 24 (2009), 3275–3281.

[13] COLLINS, ANDREW AND YAO, YIN. *Machine Learning Approaches: Data Integration for Disease Prediction and Prognosis*. 09 2018, pp. 137–141.

[14] DARRELL WHITLEY. A genetic algorithm tutorial. *Statistics and Computing 4*, 2 (1994), 65–85.

[15] DATABASE OF GENOTYPES AND PHENOTYPES (DBGAP). National institute on aging - late onset alzheimer's disease family study: Genome-wide association study for susceptibility loci, 2015. `https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000168.v2.p2&phv=76320&phd=1475&pha=&pht=707&phvf=&phdf=&phaf=&phtf=&dssp=1&consent=&temp=1`.

[16] DENG, L., AND LI, X. Machine Learning Paradigms for Speech Recognition: An Overview. *IEEE Transactions on Audio, Speech, and Language Processing 21*, 5 (2013), 1060–1089.

[17] ELLEN M. WIJSMAN, NATHAN D. PANKRATZ, YOONHA CHOI, JOSEPH H. ROTHSTEIN, KELLEY M. FABER, RONG CHENG, JOSEPH H. LEE, THOMAS D. BIRD, DAVID A. BENNETT, RAMON DIAZ-ARRASTIA, ALISON M. GOATE,MARTIN FARLOW, BERNARDINO GHETTI, ROBERT A. SWEET, TATIANA M. FOROUD, RICHARD MAYEUX, THE NIA-LOAD/NCRAD FAMILY STUDY GROUP. Genome-Wide Association of Familial Late-Onset Alzheimer's Disease Replicates BIN1 and CLU and Nominates CUGBP2 in Interaction with APOE. *PLOS Genetics 7*, 2 (2011).

[18] FOSTER PROVOST AND TOM FAWCETT. *Data Science for Business*. O'Reilly, California, 2013.

[19] GARETH JAMES, DANIELA WITTEN, TREVOR HASTIE AND ROBERT TIBSHIRANI. *An Introduction to Statistical Learning with Applications in R*. Springer, New York, 2013.

[20] GIRISH CHANDRASHEKAR AND FERAT SAHIN. A survey on feature selection methods. *Computers and Electrical Engineering 40*, 1 (2014), 16–28.

[21] HOSSIN, M. AND SULAIMAN, M.N. A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process 5*, 2 (2015), 1–11.

[22] ISABELLE GUYON AND ANDRÉ ELISSEEFF. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research 3*, 1 (2003), 1157–1182.

[23] JENNIFER WILLIAMSON ,JILL GOLDMAN AND KAREN S. MARDER. Genetic Aspects of Alzheimer Disease. *Neurologist 15*, 2 (2009), 80–86.

[24] KAROLINA SIKORSKA, EMMANUEL LESAFFRE, PATRICK FJ GROENEN AND PAUL HC EILERS. GWAS on your notebook: fast semi-parallel linear and logistic regression for genome-wide association studies. *BMC Bioinformatics 14*, 166 (2013).

[25] KELLI RYCKMAN AND SCOTT M. WILLIAMS. Calculation and Use of the Hardy-Weinberg Model in Association Studies. *Current protocols in Human Genetics, Chapter 1 Unit 1.18* (2008).

[26] KOHAVI, R., AND JOHN, G. H. Wrappers for feature subset selection. *Artificial Intelligence 97*, 1 (1997), 273–324.

[27] LARS BERTRAM AND RUDOLPH E. TANZI. Thirty years of Alzheimer's disease genetics: the implications of systematic meta-analyses. *Nature Reviews Neuroscience 9* (2008), 768–778.

[28] LYNN M. BEKRIS, CHANG-EN YU, THOMAS D. BIRD AND DEBBY W. TSUANG. Genetics of Alzheimer Disease. *Journal of Geriatric Psychiatry and Neurology 23*, 4 (2010), 213–227.

[29] M. I. JORDAN1, T. M. MITCHELL. Machine learning: Trends, perspectives, and prospects. *Science 349*, 6245 (2015), 255–260.

[30] MALGORZATA MACIUKIEWICZ, VICTORIA S. MARSHE, ANNE-CHRISTIN HAUSCHILD, JANE A. FOSTER, SUSAN ROTZINGER, JAMES L. KENNEDY, SIDNEY H. KENNEDY, DANIEL J. MÜLLERA, JOSEPH GERACI. GWAS-based machine learning approach to predict duloxetine response in major depressive disorder. *Journal of Psychiatric Research 99*, 1 (2018), 62–68.

[31] MARGARET GATZ, CHANDRA A. REYNOLDS, LAURA FRATIGLIONI, BOO JOHANSSON, JAMES A. MORTIMER, STIG BERG, AMY FISKE AND NANCY L. PEDERSEN. Role of Genes and Environments for Explaining Alzheimer Disease. *Archives of General Psychiatry 63*, 2 (2006), 168–174.

[32] MARINA SOKOLOVA, AND GUY LAPALME. A systematic analysis of performance measures for classification tasks. *Information Processing and Management 45*, 1 (2009), 427–437.

[33] MARK H. ZWEIG, AND GREGORY CAMPBELL. Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine. *Clinical Chemistry 39*, 4 (1993), 561–577.

[34] MARTIN HOFMANN-APITIUS,GORDON BALL, STEPHAN GEBEL, SHWETA BAGEWADI, BERNARD DE BONO, REINHARD SCHNEIDER, MATT PAGE, ALPHA TOM KODAMULLIL, ERFAN YOUNESI, CHRISTIAN EBELING, JESPER TEGNÉR AND LUC

CANARD. Bioinformatics Mining and Modeling Methods for the Identification of Disease Mechanisms in Neurodegenerative Disorders. *International Journal of Molecular Sciences 16* (2015), 29179–29206.

[35] MATTHEW E. STOKES, M. MICHAEL BARMADA, M. ILYAS KAMBOH AND SHYAM VISWESWARAN. The application of network label propagation to rank biomarkers in genome-wide Alzheimer's data. *BMC Genomics 15* (2014), 282.

[36] MAXWELL W. LIBBRECHT AND WILLIAM STAFFORD NOBLE. Machine learning applications in genetics and genomics. *Nature Review|Genetics 16*, 6 (2015), 321–332.

[37] MEDCALC SOFTWARE. ROC curve analysis, 2019. https://www.medcalc.org/manual/roc-curves.php.

[38] MOHAMAD SAAD AND ELLEN M. WIJSMAN. Association score testing for rare variants and binary traits in family data with shared controls. *Briefings in Bioinformatics 20*, 1 (2017), 245–253.

[39] NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. dbSNP, 2018. https://www.ncbi.nlm.nih.gov/snp/.

[40] PING ZENG, YANG ZHAO, CHENG QIAN, LIWEI ZHANG, RUYANG ZHANG, JIANWEI GOU, JIN LIU, LIYA LIU AND FENG CHEN. Statistical analysis for genome-wide association study. *The Journal of Biomedical Research 29*, 4 (2015), 285–297.

[41] QING-HAI YE, LUN-XIU QIN, M. F. P. H. J. W. K. A. C. P. R. S. Y. L. A. I. R. Y. C. Z.-C. M. Z.-Q. W. S.-L. Y. Y.-K. L. Z.-Y. T., AND WANG, X. W. Predicting hepatitis B virus–positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning. *Nature Medicine 9*, 1 (2003), 416–423.

[42] R CORE TEAM. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[43] R. MUTHUKRISHNAN AND R. ROHINI. LASSO: A Feature Selection Technique In Predictive Modeling For Machine Learning. In *2016 IEEE International Conference on Advances in Computer Applications (ICACA)* (Oct 2016), pp. 18–20.

[44] ROBERT MAKOWSKY, NICHOLAS M. PAJEWSKI, YANN C. KLIMENTIDIS, ANA I. VAZQUEZ, CHRISTINE W. DUARTE, DAVID B. ALLISON AND GUSTAVO DE LOS CAMPOS. Beyond Missing Heritability: Prediction of Complex Traits. *PLOS Genetics 7*, 4 (2011).

[45] RONALD M. NELSON, MARCIN KIERCZAK, AND ÖRJAN CARLBORG. Higher Order Interactions: Detection of Epistasis Using Machine Learning and Evolutionary Computation. *Methods in molecular biology 1019*, 1 (2013), 499–518.

[46] SANGKYU LEE, SARAH KERNS, HARRY OSTRER, BARRY ROSENSTEIN ,JOSEPH O. DEASY AND JUNG HUN OH. Machine Learning on a Genome-Wide Association Study to Predict Late Genitourinary Toxicity Following Prostate Radiotherapy. *International Journal of Radiation Oncology • Biology • Physics 101*, 1 (2018), 128–135.

[47] SEBASTIAN OKSER, TAPIO PAHIKKALA AND TERO AITTOKALLIO. Genetic variants and their interactions in disease risk prediction – machine learning and network perspectives. *BioData Mining 6*, 5 (2013).

[48] SHON, T., AND MOON, J. A hybrid machine learning approach to network anomaly detection. *Information Sciences 177*, 18 (2007), 3799–3821.

[49] STEPHEN TURNER, LOREN L. ARMSTRONG, YUKI BRADFORD, CHRISTOPHER S. CARLSON, DANA C. CRAWFORD, ANDREW T. CRENSHAW, MARIZA DE ANDRADE, KIMBERLY F. DOHENY, JONATHAN L. HAINES, GEOFFREY HAYES, GAIL JARVIK, LAN JIANG, IFTIKHAR J. KULLO, RONGLING LI, HUA LING, TERI A. MANOLIO, MARTHA MATSUMOTO, CATHERINE A. MCCARTY, ANDREW N. MCDAVID, DANIEL B. MIREL, JUSTIN E. PASCHALL, ELIZABETH W. PUGH, LUKE V. RASMUSSEN, RUSSELL A. WILKE, REBECCA L. ZUVICH, AND MARYLYN D. RITCHIE. Quality control procedures for genome-wide association studies. *Current protocols in Human Genetics, Chapter 1 Unit1.19* (2011).

[50] T. VAFEIADIS, K.I. DIAMANTARAS, G. S., AND CHATZISAVVAS, K. A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory 55*, 1 (2015), 1–9.

[51] THE 1000 GENOMES PROJECT CONSORTIUM. A map of human genome variation from population-scale sequencing. *Nature 467* (2010), 1061–1073.

[52] TIBSHIRANI, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*, 1 (1996), 267–288.

[53] VICTOR TREVINO AND FRANCESCO FALCIANI. *galgo: Genetic Algorithms for Multivariate Statistical Models from Large-Scale Functional Genomics Data*, 2018. R package version 1.4.

[54] WANG, L., ZHANG, W., LI, Q., AND ZHU, W. *AssocTests: Genetic Association Studies*, 2017. R package version 0.0-4.

[55] WILLIAM S. BUSH AND JASON H. MOORE. Chapter 11: Genome-Wide Association Studies. *PLOS: Computational Biology 8*, 12 (2012).

[56] YOONHA CHOI, ELLEN M. WIJSMAN, AND BRUCE S. WEIR. Case-control association testing in the presence of unknown relationships. *Genetic Epidemiology 33*, 8 (2009), 668–678.

[57] YU-LEN HUANG. Computer-aided Diagnosis Using Neural Networks and Support Vector Machines for Breast Ultrasonography. *Journal of Medical Ultrasound 17*, 1 (2009), 17–24.

[58] ZHI WEI, KAI WANG,HUI-QI QU, HAITAO ZHANG, JONATHAN BRADFIELD, CECILIA KIM,EDWARD FRACKLETON,CUIPING HOU, JOSEPH T. GLESSNER, ROSETTA CHIAVACCI, CHARLES STANLEY,DIMITRI MONOS, STRUAN F. A. GRANT, CONSTANTIN POLYCHRONAKOS AND HAKON HAKONARSON. From Disease Association

to Risk Assessment: An Optimistic View from Genome-Wide Association Studies on Type 1 Diabetes. *PLOS Genetics 5*, 10 (2018).

[59] Zhi Wei, Wei Wang, Jonathan Bradfield, Jin Li, Christopher Cardinale, Edward Frackelton, Cecilia Kim, Frank Mentch, Kristel Van Steen, Peter M. Visscher, Robert N. Baldassano,Hakon Hakonarson, and the International IBD Genetics Consortium. Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. *American Journal of Human Genetics 92*, 6 (2013), 1008–1012.

[60] Zhongxue Chena, Hanwen Huangb and Hon Keung Tony Ng. An improved robust association test for GWAS with multiple diseases. *Statistics and Probability Letters 91* (2014), 153–161.

[61] Ziad Obermeyer and Ezekiel J. Emanuel. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *The New England Journal of Medicine 375*, 13 (2016), 1216–1219.

Curriculum Vitae

Brissa Lizbeth Romero Rosales was born in Cuernavaca, Morelos, Mexico, on November 12, 1993. She earned the Mechatronics Engineering degree from the Instituto Tecnológico y de Estudios Superiores de Monterrey, Monterrey Campus in May 2017. She was accepted in the graduate program in Computer Sciences in July 2017. She is a committed and enthusiastic person who seeks to apply her abilities and knowledge of computer science to face complex problems mainly in health-related areas.

This document was typed in using $\LaTeX\ 2_{\varepsilon}$[1] by Brissa Lizbeth Romero Rosales.

[1] The template `MCCi-DCC-Thesis.cls` used to set up this document was prepared by the Research Group with Strategic Focus in Intelligent Systems of Tecnológico de Monterrey, Monterrey Campus.