• U  C •

FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

João Miguel Roque Almeida

Computational Discovery of Genetic Markers for Type 2 Diabetes

Thesis submitted to the
University of Coimbra for the degree of
Master in Biomedical Engineering

Supervisors:
Prof. Dr. Joel P. Arrais

Coimbra, 2018

Resumo

A Diabetes Tipo 2 é uma doença metabólica causada por resistência à insulina nos órgãos, deficiência relativa de insulina, e níveis altos de açucar no sangue. Esta é uma das doenças mais comuns no mundo, e é a quinta maior causa de morte global. Os custos estimados globais de tratamento tanto directo como indirecto, chegam a atingir os US$1.31 trillion (95% CI 1.28 - 1.36). Como tal, torna-se cada vez mais important descobrir métodos que possam prever o risco da DT2 desde uma idade jovem, e sem que até nenhuns padrões de risco fisiológicos se verifiquem. Com isto, será possível tanto para médicos como para pacientes estar mais conscientes do risco da doença e poderem empregar medidas preventivas o mais cedo possível.

Existem indícios claros que apontam a Diabetes Tipo 2 como uma patologia influenciada não só por factores ambientais, mas também genéticos. Por isso, este estudo pretende desenvolver novas abordagens a *Genome Wide Association Studies*, mais especificamente no que trata a análises Multi-Locus em doenças complexas, que sejam não só computacionalmente praticáveis mas que estudem também a não-linearidade nestes tipos de dados. Para o fazer, foi desenvolvida uma nova linha inovadora de transformações que permite identificar regiões de interesse no genoma, extrair novas características sem perder o contexto biológico do problema, e utilizá-las em modelos de *Machine Learning* que acontam com a epistasia.

Estes novos métodos são demonstrados numa análise de um dataset de Polimorfismos de Nucleótidos Únicos, onde novos possíveis marcadores genéticos para a Diabetes Tipo 2 são apontados. Para além disso, também é realizada uma classificação do risco de DT2, com *F1-Scores* a atingir os 0.97 com alta confiança. Este projecto pretende sobretudo mostrar como podem ser minados os dados de *datasets* de genótipos de uma maneira que permita o uso de modelos de *Machine Learning* com a sua capacidade total.

Palavras-Chave: Aprendizagem Máquina, Diabetes Tipo 2, Estudos de Associação Genética, Bioinformática, Análise de Dados, Genética.

Abstract

Type 2 Diabetes is a metabolic disorder caused by insulin resistance in organs, relative insulin deficiency and high blood sugar levels. It is one of the most common diseases in the world, and the fifth leading cause of death worldwide, with an estimate global cost of indirect and direct treating reaching US$1.31 trillion (95% CI 1.28 - 1.36). As such, it becomes increasingly important to discover methods of predicting T2D risk from a young age and before the onset of any physiological risk patterns, so that both patients and doctors are aware of it, and can monitor the disease and employ preventive measures.

There is clear evidence that supports Type 2 Diabetes risk as being influenced not only by environmental factors, but also genetic ones. In light of this, the following study aims to develop new ways to approach Multi-Locus Genome Wide Association Studies in complex diseases, that are not only computationally feasible, but can also study the non-linearity in a dataset. It aims to do so through the inclusion of an innovative pipeline of transformations that can identify regions of interest in the genome, extract new features without losing biological context of the problem and use them in Machine Learning models that can account for epistasis.

This process is further demonstrated in an analysis of a Single Nucleotide Polymorphisms dataset, and provides several identifications of possible novel genetic markers for Type 2 Diabetes. Furthermore, classification of T2D's risk is also performed, reaching F1-scores as high as 0.97 with high confidence. This project aims mostly to exhibit how can a genotypes dataset be data mined in a way that can be fully taken advantage of by Machine Learning models.

Keywords: Machine Learning, Type 2 Diabetes, Genome Wide Association Study, Bioinformatics, Data Analysis, Genetics.

Acknowledgments

First of all, I'd like to thank my advisor Joel P. Arrais, for all the support he gave me throughout this project, despite all the incredible things that were happening in his life, such as the birth of son, to whom I give my best wishes. He really did handle me a great deal of responsibility and liberty to investigate what I truly wanted, in the subject that I wanted, and for that I am extremely thankful.

I want to thank Conceição Ega and Hugo Froufe from UC-BIOTECH that kindly shared the dataset used in this project with me, and helped answering so many of the questions I had as of the start of this project.

Junior Enterprise for Science and Technology was one of the most surprisingly positive experiences I had the pleasure to be a part of during my time in this University. I want to thank for the incessant discussion of the most random subjects one can think of, for all the shared knowledge, for letting me feel truly welcome as I am, and above all, for being such a bunch of amazing people that keeps growing every year.

To all my friends for the company through these years, for all the laughs, shared moments, work and all nighters. Thank you for being there, sharing joy and pain, for being you. I do really feel like I have a great number of true friends, who have proven so countless times, and I hope to return to you, as much as you give to me.

Wee Ana, there is simply no way of putting on paper how much your presence as positively impacted my life. Thank you for putting up with me, you really do rock my world.

Finally to my father, my mother, my brothers, my uncle and my grandmother, for giving your all in these last tough times, and still having the strength left to be so present as you are in my life. I am really proud to be able to call you family.

"*A scientific man ought to have no wishes, no affections, - a mere heart of stone.*"

Charles Darwin

Contents

List of Tables

List of Figures

Abbreviations

AUC	Area Under Curve
BMI	Body Mass Index
CNV	Copy Number Variation
DNA	Deoxyribonucleic Acid
DNA	Next Generation Sequencing
DNN	Deep Neural Networks
GDP	Gross Domestic Product
GWAS	Genome Wide Association Studies
HWE	Hardy-Weinberg Equilibrium
INDEL	Insertions or Deletions in the genome
LADA	Latent Autoimmune Diabetes of Adulthood
LD	Linkage Disequilibrium
LDA	Linear Discriminant Analysis
MDR	Multifactor Dimensionality Reduction
OR	Odd's ratio
PCA	Principal Component Analysis

QTL	Quantitative Trait Loci
RBF	Radial Basis Function
ROC	Receiver Operating Characteristic
SNP	Single Nucleotide Polymorphism
SV	Structural Variants
SVM	Support Vector Machine
T1D	Type 1 Diabetes
T2D	Type 2 Diabetes
VCF	Variant Call Format
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing

1

Introduction

1.1 Context

1.1.1 Genetic Markers

The human genome is composed of around 3 billion nucleotides, them being A, C, T and G, adenine, cytosine, thymine and guanine respectively, with 23 pairs of chromosomes, 1 of them being the sexual chromosomes [9]. It is known that individuals from the same population have similar DNA than to those of different ones [10]. The small differences between the genomes can be Single Nuclear Polymorphisms, Mutations, Insertions, Deletions and Copy Number Variations, meaning that the variants mentioned are known to be responsible for most of the different phenotypes (observable characteristics) in humans [11]. Mutations are extremely rare alterations on the DNA, Indels are, as the name indicates, insertions or deletions that may or may not occur from individual to individual and CNVs are sections of the genome that are repeated, and the number of repetitions varies between individuals [11].

SNPs are single nucleotide variations that occur in a specific position in the genome, typically with two alleles, and they can either be rare or common. There are around 10 million SNPs in the human genome, meaning that on average a SNP occurs every 300 base pairs [11]. Allele frequencies are given for the most common one, for example, if a SNP can either be a T or a C, if T is the most common one with 0.7 allele frequency, 70% of the population will have a T, and the rest a C [7]. Common variants are the ones with a minor allele frequency of 5% or more, and rare variants are only present on less than 5% of the population. All of these genetic variants can be used as markers to find associations between genotypes and phenotypes, with the most commonly used for the effect being SNPs, due to their abundance. This association can be fairly straightforward in case of single gene related traits or diseases, but not so much for more complex traits, such as the case of T2D [9].

1.1.2 Type 2 Diabetes and Missing Heritability

Type 2 Diabetes is a metabolic disorder caused by insulin resistance in organs, relative insulin deficiency and high blood sugar levels [12]. There is evidence supporting that T2D is strongly influenced by genetic and epigenetic factors, as well as environmental ones. Throughout their lifetime, individuals with one parent who has T2D, have a 40% risk of developing T2D. This risk increases to 70%, if both of the parents have T2D [1, 13]. Studies performed with twin pairs show a lower discordance rate (one of the twins has the disease, and the other doesn't) for monozygotic twins than for dizygotic, which supports the genetic and epigenetic influence on T2D even further [14]. Furthermore, variants associated with T2D in the European population might not be replicated in other non-European populations, and vice-versa. A higher prevalence of the disease is also seen in some populations [15, 13, 16]. However, our genetic code doesn't undertake significant alterations in only one or two generations, so this recent surge in predisposition for T2D is also due to the gene-environment interactions. Increase of adipose tissues in human populations is the single most significant factor in this epidemic, and to model the interaction between genes and causes that lead to obesity is extremely complex. Who burns more calories at rest, who has greater exercise levels when not doing it actively, who is more willing to change a sedentary lifestyle are all examples of gene influencing behaviour, and that's what makes the gene-environment interaction so hard to include in Genome Wide Association Studies [1]. It is also important to note that there is no formal definition for T2D, since all the cases who do not fulfil the criteria for T1D, LADA and other types, are considered T2D. It's a disease more associated with age, although it also has been reported in adolescents [17]. The question of how to clinically phenotype T2D is very important, because it can influence its association with genotype, since providing different patterns for the same phenotype will make it harder to perform classification [15].

So far, for T2D, more than 80 robust markers were found, even though they only account for 20% of the heritability for this disease. Even more so, these markers are predominantly common, with additive effects [8]. The hypothesis that a common disease can be caused by several common variants in the genome is not new, and several studies have already identified common alleles who play a role in certain traits or disease susceptibility [18, 8, 7, 19]. The remaining hypothesis are that a few rare variants have big effects (common disease-rare variant), and that both common and rare variants play a part in susceptibility [13, 15]. Despite the successes of GWAS in identifying markers, much of the heritability in complex diseases still

remains unexplained, which leads us to the missing heritability problem.
Heritability is defined as a ratio:

$$\pi_{explained} = h^2_{known}/h^2_{all} \tag{1.1}$$

, where h^2_{known} is the proportion of the phenotype explained by known variants that affect it, and h^2_{all} all the variants, including those who remain undiscovered. The underlying problem is that the h^2_{all} might not be properly estimated, which leads to an underestimation of $\pi_{explained}$. This model also fails to consider epistasis, that can greatly inflate the apparent heritability, and it is not yet consistently detected with the current standard methods available [20]. It is important to acknowledge the missing heritability problem, because the reasons why it might be happening are related to how GWAS are thought-out and performed. The first reason is that common variants of low frequency (1-10%) might not be identified because of the genotyping arrays themselves lacking useful proxies. Secondly, many common variants with very small effects can be extremely hard to identify with current sample sizes [21, 22].

1.2 Motivation

Type 2 Diabetes is one of the most common diseases encountered in the world, and the fifth leading cause of death worldwide. Data from the International Diabetes Federation has shown that, in 2011 there were 366 million people in the world living with diabetes, and that number is expected to rise to 552 million by 2030, 80% to 90% of the cases being of T2D [15, 13]. In 2015, diabetes caused 5 million deaths worldwide, with an estimate global cost of indirect and direct treating of US$1.31 trillion (95% CI 1.28 - 1.36) [23]. In Portugal alone, the costs related to T2D corresponds to 1% of the country's GDP [24].

Type 2 Diabetes risk factors, regardless of ethnicity or genetic risk, are elevated fasting insulin concentrations and low insulin secretion, obesity and fat distribution, caused by poor diet, lack of physical exercise and smoking [25, 26]. It has also been shown that changes on this behaviour at an individual level, for a more supportive environment and healthy lifestyle can greatly delay or prevent entirely Type 2 Diabetes [27, 28].

The first time a whole human genome was sequenced in 2001 it cost around US$300 million [29]. Since then, the aim has been to reduce it to US$1000 per genome, and

so far, that goal is very close to being reached [30]. In a span of a few years, the general cost of genome sequencing decreased immensely, which lead to an increase in the number of Genome Wide Association Studies performed. As such, numerous regions of Linkage Disequilibrium in the genome that are associated with certain traits or diseases were discovered, which makes possible to identify an individual's elevated risk for certain genetic diseases [31].

Despite all the efforts, a convincing T2D risk predictor has not yet been attained [16]. Such discovery would be a huge step in respect to personal healthcare, since from birth, doctors and patients would be more aware of certain disorder risks. By discovering more meaningful genetic markers for T2D, and by finding new ways to analyse them in the genome, it should be possible do develop a risk predictor that can be used to better inform both doctors and patients. This would hopefully lead to a much earlier prevention and monitoring of the disease, even before any physiological signs are present. [32]

1.3 Objectives

The main goal of this thesis is to develop a Type 2 Diabetes predictor from a Single Nucleotide Polymorphisms dataset, that is able to return information of important variants that are relevant to the problem. Since only data from the Iberian Peninsula is being used, it is of extreme relevance to establish a method that can be replicated on other ethnicities. To do so, there are five essential objectives to be fulfilled:

1. Prepare a complete dataset, with the most possible correct and corresponding variants for cases and controls, that enables an accessible investigation of said variants and makes possible the application of machine learning.

2. Develop a pipeline of feature engineering that can be replicated for any dataset of SNPs, without losing their biological context.

3. Discover novel possible markers and verify the presence and impact of already known genes that increase T2D's susceptibility.

4. Through the use of Machine Learning models, implement a T2D risk predictor.

5. Build a model that can be further validated in future work when more data is available.

1.4 Structure

The chapter State of the Art, performs a quick showdown of the technologies used since the translation of DNA to analysable files, and the most current practices when doing such analysis. It also goes further into some Machine Learning and Data Mining methods that are currently being used with great success in the most diverse areas, and how some of them are applied to GWAS.

The third chapter, Data Preparation, describes the processes used to prepare and clean all the raw data into formats that can be used to perform Machine Learning. Successively, the Feature Engineering chapter explains the methods that are employed to find new variants, and reduce the number of available features into a smaller and more relevant subset without losing information.

The fifth chapter characterizes how the Machine Learning techniques were used in this context and shows the following results. Lastly, there are the Discussion and Conclusion, where a description of the whole process is discussed, as well as advantages and disadvantages of it. Future approaches and the overall place for Machine Learning in GWAS are also addressed.

2

State of the Art

2.1 Genome Sequencing

Over the last years of genome sequencing innovation, a shift has been seen from the more traditional Automated Sanger sequencing to cheaper and faster Next Generation Sequencing techniques. For a quick overview, the Sanger method starts of with either bacterial cloning or a Polymerase Chain Reaction to amplify the DNA strands. It is followed by four reactions containing deoxynucleotides and DNA Polymerase performed separately for each to include a different dideoxynucleotide (ddNTP). When a ddNTP that binds to a specific deoxynucleotide is attached to an elongating DNA chain, the DNA polymerase stops it's process on it, thus labelling every nucleotide. The ddNTPs are radioactively or fluorescently marked so that the final sequence can be visualized on the gel electrophoresis image, such as in the figure 2.1 [2].

The most recent implementations of this method can achieve extremely high genotyping accuracies, but are expensive, time consuming, and better suited for only single gene sequencing. Since there was a need for cheaper processes with larger throughput capabilities, Next Generation Sequencing emerged. The single major advantage provided by this new technology it's is ability to output enormous amounts of data in a single run, cheaply and fast. DNA methods start off by randomly fragmenting the DNA, and coupling adapters to both ends of the fragments. These are then attached to a flow cell and amplified in clusters that are recorded by various pictures. This produces a great number of short reads that are then anchored to a reference genome, where small differences can be determined. The number of base pairs contained in every read, and the number of reads depend on the sequencer itself [33]. Further description and visualization of this process can be found in the figure 2.2.

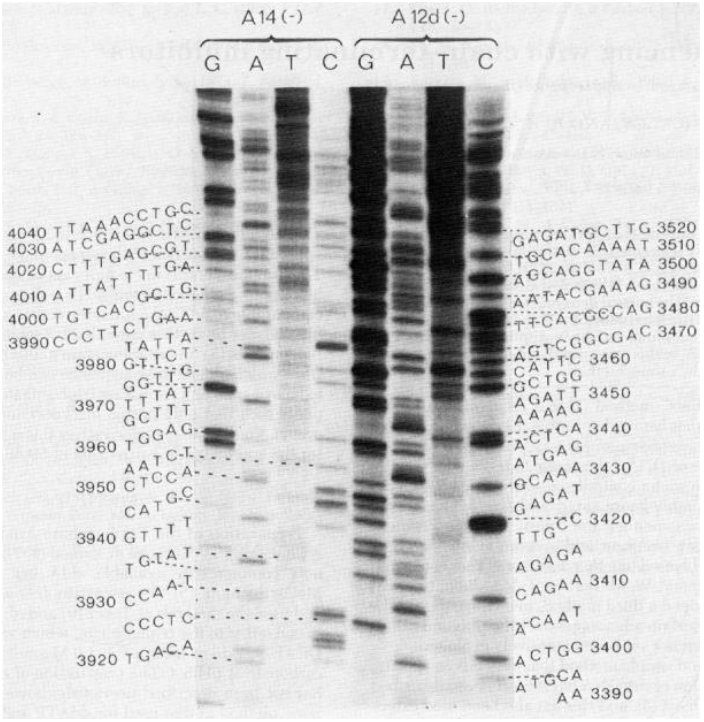

Figure 2.1: Picture of the acrylamide gel electrophoresis with the sequences read annotated on the sides. From left to right, the inhibitors used are ddGTP, ddATP, ddTTP, and araCTP. Adapted from "DNA sequencing with chain-terminating inhibitors" [2].

Generally, genome studies only target the exome because it contains the protein coding regions [34]. However, it has also been shown that many variants that are associated with disease can be found on non coding regions of the genome [35]. For this reason, it became increasingly important to integrate WGS techniques when performing GWAS. Nonetheless, the cases data utilized in this study are only of the exome.

From the data recovered, it is then possible by using reference panels of known SNPs or other variants to produce VCF files, that record the genotype for each variant [36, 37]. There are several programs available to perform variant calling from the short reads aligned to a reference genome, such as SAMtools (*http* : *//samtools.sourceforge.net/*). These can do this process in many different ways, but the main goal is to identify where the reads differ from the reference genome and to translate those to a VCF file.

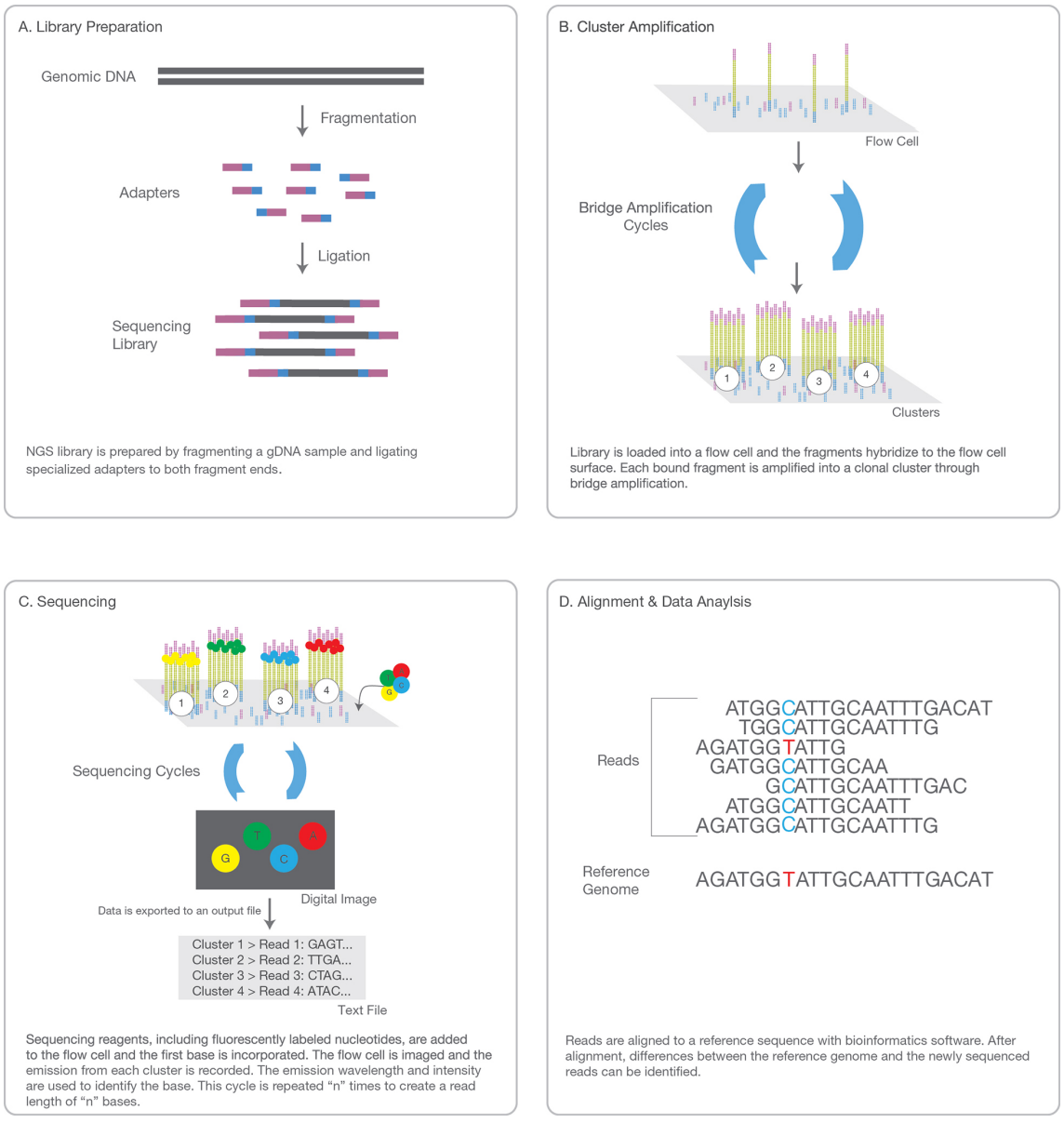

Figure 2.2: Work flow of the Next Generation Sequencing Techniques employed by Illumina sequencers. Adapted from Illumina's images for general use.

2.2 Variant Call Format

Nowadays, as previously stated, information of the genotypes assembled from the sequencers are represented in a Variant Call Format files, or VCF files. These start with the meta information lines, which indicate the VCF version, followed by the three possible parameters INFO, FILTER and FORMAT. INFO describes any additional information of the metrics collected for every variant and FILTER describes what kind of filters were applied to them. FORMAT indicates the type of genotyping information that is present and it's characteristics. Then, the header line ensues which presents the following 8 fixed columns: #CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO. These represent respectively the chromosome of the variant, it's position in it, a sample ID, the bases for the reference genome, the bases for the possible alternative bases for the variant, the quality or certainty of the genotype, any filters applied that were previously described and any additional info for the variant. If there is genotype information present, then a FORMAT columns ensues, which dictates how the genotypes are going to be presented, followed by the actual genotypes for a set number of samples. All the following lines represent existing variants that were genotyped. More information on this file type can be found at $https://github.com/samtools/hts-specs$.

2.3 Single Marker and Multi-locus Analysis and Imputation

Since 2007, GWAS have been responsible for the discovery of genetic markers that relate to complex human traits and disorders, the simplest approach being the Single Marker Analysis, which is both capable of finding common variants with addictive effects and rare ones with high phenotype impact [38].

This marker-by-marker analysis focus on the individual effect of each variant, to detect associations between molecular markers and traits or disorders in a population. These close groups of variants that are discovered to be correlated to a certain phenotype are called Quantitative Trait Loci. In many complex diseases, several QTLs are discovered because of Linkage Disequilibrium, that refers to the non-random association of alleles at different loci in a population, and it's visible when their association frequency is higher or lower than what would be expected if it was in fact random [39]. It is influenced by many factors, such as rate of muta-

tion, population structure, genetic linkage, the rate of mutation, genetic drift and the system of mating, and can signal segments in the chromosomes that trace back to a common ancestor without intervening recombination. Essentially, the objective is to determine if phenotype differences are due to a few loci with large effects or many loci with small, but additive ones.

We can formalize a contingency table that contains the information of the genotype for each SNP, and the corresponding phenotype for each sample, and use tests such as **Pearson's χ^2 Test**, **Fisher's Test** and **G-Test** to detect an association between allele frequencies and phenotype [9]. Since the **Pearson's χ^2 Test** is going to be used, a short description follows. Normally, this test is used to verify if there is a deviance of the expect frequencies and the observed frequencies. It's null hypothesis is that the data are independent, and by rejecting this hypothesis, we can then find associations between them. This test assumes the data follows a normal distribution, and it's given by [40]:

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - m_i)^2}{m_i} \tag{2.1}$$

To go even further and detect a QTL with this approach, the association between a marker and a trait, or in this particular case, Type 2 Diabetes, can be modelled by simple regression methods [41]. The null hypothesis for these tests is that the marker has no association with the trait, and to test it, a t-test, F-test or Bayes factor can be used. We assume that a marker will only affect the desired trait if it is in LD with a QTL. To measure LD, considering A and B as two different markers, A1, A2, and B1, B2 as their alleles respectively:

$$D = frequency(A_1\_B_1) \times frequency(A_2\_B_2) - frequency(A_1\_B_2) \times frequency(A_2\_B_1) \tag{2.2}$$

, $frequency(A_1\_B_1)$ being the frequency of the haplotype $A_1\_B_1$, and likewise for the remaining haplotypes. Since this is very dependant on allele frequencies, the r^2 metric was proposed:

$$r^2 = \frac{D^2}{frequency(A_1) \times frequency(A_2) \times frequency(B_1) \times frequency(B_2)} \tag{2.3}$$

By combining r^2, LD and the number of markers in the study we can infer the power of the association test to detect QTLs [42, 43].

The following approach tried was the Multi-Locus analysis. It was the next logical

step, because it tries to tackle issues that were not handled in the previous methods, such as epistasis and big marker spacing (less genotyped variants) [44]. It is fairly recent, and its objective is to consider various locus and their interactions when performing the association studies. When building marker-to-marker association models, if our set contains 1 million SNPs, it becomes both statistically and computationally hard to differentiate between significant markers and to understand their biological context, as there will be 5×10^{11} interactions to examine [45]. Although a Multi-Locus analysis adds several benefits to it's previous iteration, it requires even more processing power. At this point, feature selection comes into play, there being several ways to approach it. The most common ones are the usage of SNPs that have met certain criteria in broader previous tests or integrating biological knowledge in the models. Of course, either one of these strategies imposes bias, such as eliminating potential markers that don't prove to have a significant effect on their own, and missing novel undocumented interactions [46]. By considering a smaller number of SNPs in which the phenotype information might be contained, it becomes possible to make such an analysis. One example of this approach is the Biofilter, that uses previous Biological knowledge to construct several multi-SNP models, and only then applies Logistic Regression and Multi-factor Dimensionality Reduction methods to perform its analysis [7, 47].

To increase the number of SNPs available and generate a common set of genotyped variants for each dataset, for Genome Wide Association Studies, Imputation surfaced as a viable candidate. By using known Linkage Disequilibrium patterns and frequencies of haplotypes from the 1000 Genomes Project, it is possible to make a correct estimate of missing genotypes. However, imputation leads to the underlying assumption that the study population has the same patterns of LD and that the association between haplotypes and causal loci is the same in the reference population, which might not always be the case [48, 7].

2.4 *Bayesian* Methods

The usage of previously explained *frequentist* methods, although very widely used, still pose some problems because of a limitation on the usage of *p-values* themselves. From a *p-value* alone it is very hard to know how confident it is possible to be when a SNP is associated with a phenotype [49]. Furthermore, the datasets used in these studies are usually small, an it is necessary to account for their uncertainty [50]. By using *Bayesian* methods we can circumvent these issues, at the expense of addi-

tional assumptions about the influence on phenotype for each SNP. Another great advantage of these methods is that they allow for a common-ground when comparing results between studies (*meta-analysis*), facilitating knowledge integration [51]. It turns the probability that a SNP affects a phenotype in a quantitative measure, the Bayes Factor:

$$BF = \frac{P(data|\theta_{het} = t_1, \theta_{hom} = t_2)}{P(data|\theta_{het} = 0, \theta_{hom} = 0)} \tag{2.4}$$

For these reasons, *Bayesian* methods have become more prevalent in recent years in GWAS [52]. It is also readily usable in several packages, such as SNPTEST [53], genMOSS [54] and BIMBAM [55].

2.5 Dimensionality Reduction

However, even with such techniques, the missing heritability is still yet uncovered for complex diseases, which leads us to Epistasis or gene-gene interaction, and how to integrate it in GWAS. These are some of the methods available that can enable Multi Locus analysis. When accounting for gene-gene interactions, as stated before, the problem becomes statistically and computationally complex, since for the three possible genotypes and k SNPs, there are 3^k genotype classes possible [56]. To attenuate these issues, and combine variants information considering gene-gene interaction, Dimensionality Reduction can be used. Usually these problems entertain very large numbers of SNPs so combining their information in a smaller number of vectors while still retaining most of their variance information can make it extremely easier to analyse these kinds of datasets. There are several methods that can perform this reduction, such as MDR, PCA and LDA.

At it's core, MDR is an algorithm capable of constructing new features by pooling genotypes from multiple SNPs [57]. Considering several multi locus genotype information and given a threshold T, a certain group of SNPs is considered high risk if their ratio of case study to control group is higher than T, or low risk if that same ratio is lower [4]. By doing so, a new one dimension vector with two different groups (High and Low risk) is constructed, which enables the usage of other techniques to process it, and produce multi locus analysis results.

The PCA's goals are to extract a set of new orthogonal variables of a dataset, called the Principal Components, that are able to represent the important information in a k number of vectors [58]. To do so, a p-dimensional vector of weights $W_{(k)} =$

$(w_1, w2, ..., w_p)_{(k)}$ that map to each row of our dataset matrix X will be worked out. To calculate the first component we maximize the variance:

$$W_{(1)} = arg\ max\Big\{\frac{W^T X^T X W}{W^T W}\Big\} \tag{2.5}$$

, which is the largest eigenvalue of X when W is the corresponding eigenvector. The first Principal Component will then be given by $t_{1i} = X_{(i)} \cdot W_{(1)}$ [59]. Considering n observations and p features, this method can output several $\min n-1, p$ vectors containing the highest variance of the data possible. This method is usually used in GWAS to detect population structure and outliers [60].

The last dimensionality reduction method covered is Linear Discriminant Analysis and it makes use of the data labels to propose a linear combination of variables that best maximize the classes separation. It is then, contrary to PCA, called a supervised algorithm. Regarding a binary problem, which happens to be most of phenotypic studies, for each class in y, mean and covariance are represented by μ_0 / $\sum_0$ and μ_1 / $\sum_1$ respectively. A function of the linear combination of samples can be obtained by:

$$\overrightarrow{w} \cdot \overrightarrow{x} > c \tag{2.6}$$

where,

$$\overrightarrow{w} = \sum^{-1}(\overrightarrow{\mu_1} - \overrightarrow{\mu_0}) \tag{2.7}$$

$$c = \frac{1}{2}(T - \overrightarrow{\mu_0}^T \sum_0^{-1} \overrightarrow{\mu}_0 + \overrightarrow{\mu_1}^T \sum_1^{-1} \overrightarrow{\mu_1}) \tag{2.8}$$

and T is a threshold that verifies the following condition:

$$(\overrightarrow{x} - \overrightarrow{\mu_0})^T \sum_0^{-1}(\overrightarrow{x} - \overrightarrow{\mu_0}) + \ln|\sum_0| - (\overrightarrow{x} - \overrightarrow{\mu_1})^T \sum_1^{-1}(\overrightarrow{x} - \overrightarrow{\mu_1}) - \ln|\sum_1| > T \tag{2.9}$$

The new hyperplane defined by c, is then the one that maximizes the separability of classes. It can be used either for classification or to reduce dimensionality of a dataset [61].

These approaches have been widely and successfully used in other GWAS studies, and improved with entropy-based interpretation methods, the use of odds ratio, imputation, parallel implementations and much more [57]. Furthermore, cases like the susceptibility to bladder cancer were found in highly significant interactions between SNPs using such methods that consider epistasis, with higher prediction

power than smoking [62].

2.6 Quality Control and Validation

When performing a GWAS, the ultimate goal is for it to be able to predict, in any new given dataset, the interaction of markers and phenotypes that were discovered. To be able to validate this work, it is extremely important to test the results in data sets that were not used during the study, since most significant effects uncovered are likely to be overestimations [63]. In GWAS, there are far more SNPs than number of samples, which can easily lead to a model that predicts all cases in the discovery dataset and that cannot be replicated in others. This is called overfitting [64]. Furthermore, validation leads to a higher confidence level in the study performed, and also allows to uncover the populations where it can be replicated, if at all. Validation must be first thought out when designing the study, by assigning a percentage of samples to perform testing and other for cross-validation (apart from training).

As most of the studies samples are from a single population, the choice of control group must also take that into consideration as a risk variant might not be relevant across all different populations. Most control groups can be gathered from the study itself, but to promote less bias, samples from the 1000 Genome Project can be used [48]. This project gathered variants from populations across the world, that serve not only as control, but as gold standards for imputation.

The verify the quality of the genotypes themselves, it is possible to use the Hardy-Weinberg Equilibrium [65]. The HWE is a model that states genotype frequencies follow certain rules, and remain constant at each generation [66]. Some factors that can affect this equilibrium include migration, mutation, natural selection and assortative mating (tendency for people to choose partners who are more phenotypically similar or dissimilar to themselves). However, if these factors are negligible in the target population, deviations from the HWE are most likely due to incorrect genotyping. Considering $f(A) = p$ and $f(a) = q$, the expected genotypes frequencies are then:

$$f(AA) = p^2 \tag{2.10}$$

$$f(aa) = q^2 \tag{2.11}$$

$$f(Aa) = 2pq \tag{2.12}$$

As such, the HWE should follow the curved line in the De Finetti diagram displayed in figure 2.3.

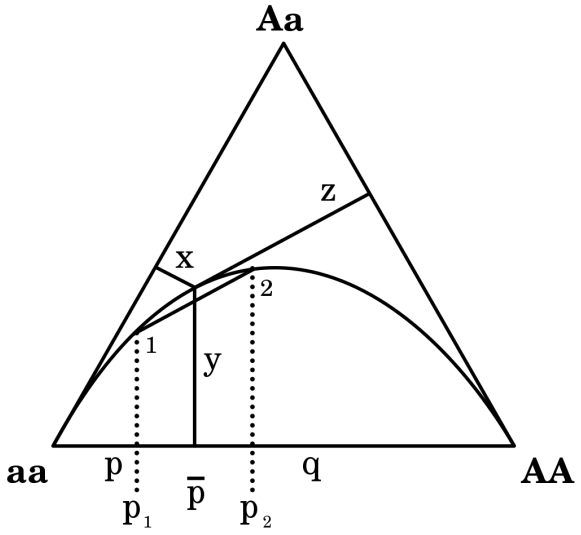

Figure 2.3: De Finneti diagram of genotype frequencies. If a population follows the Hardy-Weinberg Equilibrium, their genotype frequencies distribution will follow the curved line on the plot. Free licensed image from Wikimedia Commons.

2.7 Data Mining and Machine Learning

In genomics, the vast quantities of data that must be scouted before any meaningful results are achieved is daunting, and simple frequentist methods have not yet been able to fully crack the problem of complex diseases heritability. As we inquire further the study of gene-gene or gene-environment interactions and non-linearity in the mapping of genotype to phenotype to understand genomic variation, disease susceptibility and the role of environment in genomics, it is important to evaluate how these situations are approached. There might be a combination of SNPs that, if addressed with the proper non-linear function, can significantly translate into the respective phenotypes, but each SNP individual contribution might not appear any different than the millions of other SNPs. To this outcome, that can't be predicted by the sum of all markers, it is called non-linearity [4].

To produce non-linear models, it first must be discussed how to effectively reduce the amount of SNPs that a model needs to look through, or in other words, how to

data mine the genome. At this point, the scientific community is slowly transitioning to non-linear analysis of the genome, but there aren't many methods developed so far [67]. A model can also be built without any preprocessing, but besides the great quantity of features for few samples, a great majority of the millions of SNPs available are considered either noise or non-significant to the task at hands. There are essentially 2 types of feature selection used, which are the filtering and the wrapper approaches [4]. The first one, refers to the preprocessing of data and assessment of the quality and significance for each feature, to ultimately collect a significant subset. The wrapper one utilizes a deterministic or stochastic algorithm that iteratively selects subsets of data to classify. Filtering is a faster approach, while the wrapper can be more powerful, since it doesn't discard variables with assumptions of quality. Even if these two methods are widely used, there are multiple other ways of discarding noisy SNPs, for example, the inclusion of biological knowledge [68].

After the preprocessing methods are carefully selected and implemented, and subsetting has been performed, machine learning can be used to classify between the desired phenotypes. The purpose of machine learning is to make computers learn how to perform certain tasks, by providing them with a set of examples, but without explicitly programming them to do so. In our problem context, each variant is considered a feature, which means that it is an attribute from where the model can learn. The entirety of features and samples compose the training dataset. To provide the machine with the context of what to learn, a target vector is given. This vector contains information on the phenotype that requires classification [69]. The most popular machine learning methods are Support Vector Machines, Decision Trees, Naïve Bayes classifiers, Neural Networks and Fuzzy Sets [70]. Since there is a growing interest in non-linear models, SVMs, Decision and Neural Networks are some of the most promising Machine Learning Methods [71]. On this project, SVMs and Decision Trees were used, so a more complete description is provided for them.

For explanation purposes, let's consider the binary classification problem, with linearly separable classes and only two features for the SVM. A Support Vector Machine tries to ensure the largest possible boundary between both classes. To do this, it produces two support vectors to the main decision boundary, that provide the largest distance possible between the closest instances from each class [72]. This process can be better understood through figure 2.4. When more features are added to the problem, it stops being viewable in plots, since the dimensionality of the problem grows, but the same principle is applied.

However, this mechanism assumes that there are no instances overlapping which

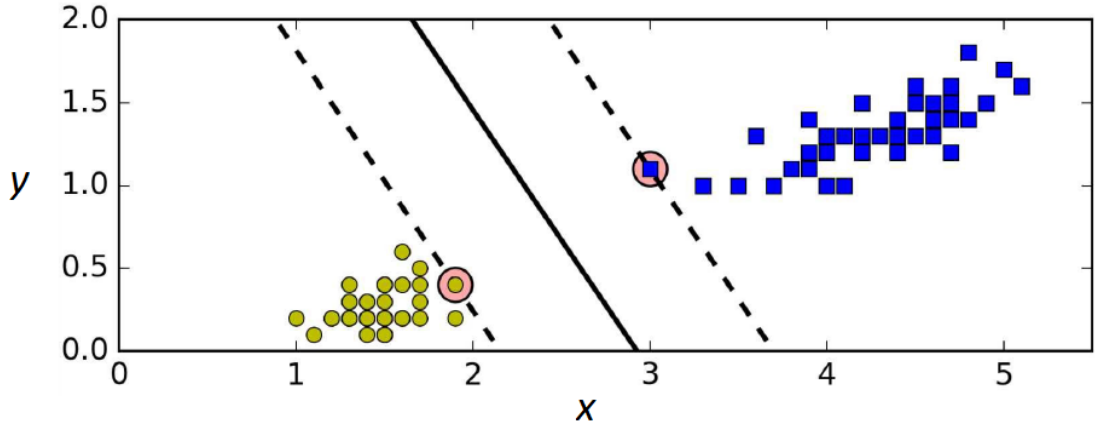

Figure 2.4: Plot of two classes separated by the support vectors on the dashed lines, and the decision boundary on the black line at the centre. Adapted from the book "Hands on Machine Learning with Scikit-Learn and TensorFlow" [3].

is something very likely to occur in a "real-life" dataset. To then handle it, a soft margin classification can be implemented. What it does, is allow for some violations of the margins that support vectors provide, which makes for a wider distance between them. By doing this, it is also much more likely that the classifier will generalize better. The parameter that handles the number of violations allowed is C [3]. The higher it is, the fewer the violations and consequent distance of support vectors. An example of $C = 1$ can be seen in figure 2.5.

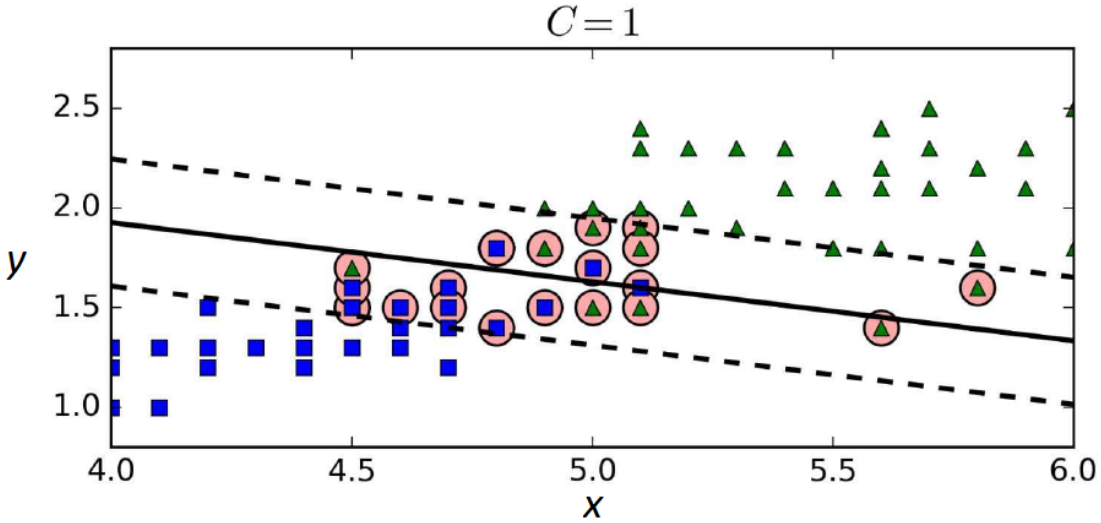

Figure 2.5: Plot of SVM adjusting to two features with $C = 1$. Adapted from the book "Hands on Machine Learning with Scikit-Learn and TensorFlow" [3].

Although this classifier is very powerful, a very large number of datasets cannot be linearly separable. To solve this, extra features that are transformations of the

original ones are added, that allows for a different arrangement of data. An example of this process can be seen in figure 2.6. However, performing such great number of transformations can turn extremely computationally heavy which lead to the implementation of the *Kernel Trick*. This trick is intended to calculate the dot product of the transformed vectors without having to transform them [3].

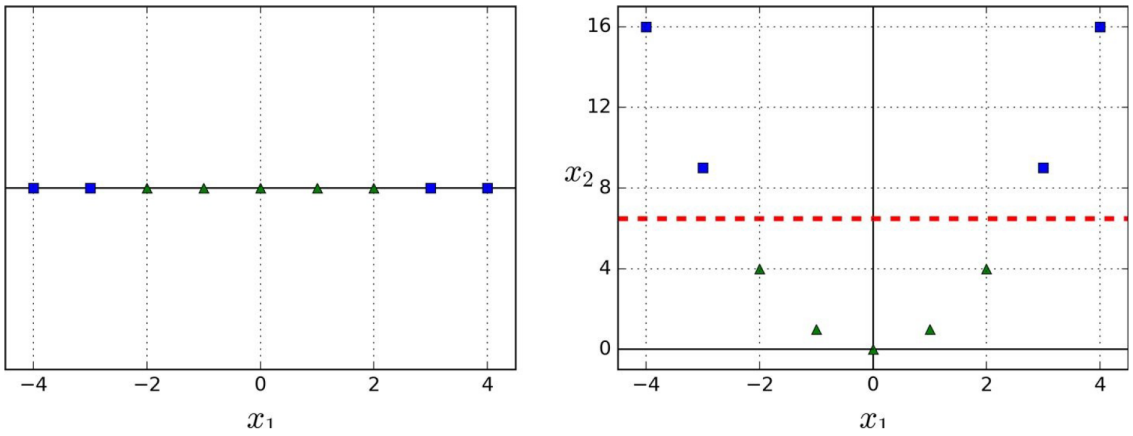

Figure 2.6: Demonstration of features transformation $x_2 = (x_1^2)$ to find non-linear relationships. Adapted from the book "Hands on Machine Learning with Scikit-Learn and TensorFlow" [3].

So in Machine Learning, a kernel is a function capable of calculating the dot product of the transformed vectors based on the original ones. Some common kernels are Linear, Gaussian Radial Basis Function, Polynomial and Sigmoid.

The following method described is Decision Trees. Decisions Trees, the unitary block of Extra-Trees, are a group of sequential if-then-else rules (nodes) that break down the dataset in ever so smaller subsets. These rules, are based of a single feature and threshold, which try to look for the combination that splits the data into the purest subsets. This keeps happening recursively, until it reaches the maximum depth of the tree, or until it can no longer reduce impurity. There are several criterion used to measure impurity, some of them being the Gini impurity or entropy [4, 73].

To improve on their classification power, methods which are ensembles of Decision Trees were developed. These are entire "forests" built using trees as unitary blocks, the most used being Random Forests and Extremely Randomized Trees classifiers. The Extra-Trees classifier, makes use of a set number of Decision Trees, where at each node, only a subset of random features is considered. Also, rather than searching for the best possible thresholds, it utilizes random ones too. Then, to make predictions, all the votes from every single tree are counted to output a final decision. Feature importances are calculated according to their depth in each tree.

Features with high purity are usually the first ones being selected for a rule, which makes their importance score higher [74]. This exact process can be seen on figure 2.7. Random Forests work in a very similar way, but the best possible thresholds are calculated, which makes them more computationally heavier than the former, although slightly more accurate [75]. Both methods are easily interpretable and applicable in case-control studies, and are highly adaptive to data, which makes them effective when dealing with "large p, small n" problems. Besides these advantages, they also account for interaction between variables, making them tailored to detect epistasis [76]. These classifiers can be used to perform SNPs selection, genotype-phenotype association, epistasis detection and risk assessment [77, 78].

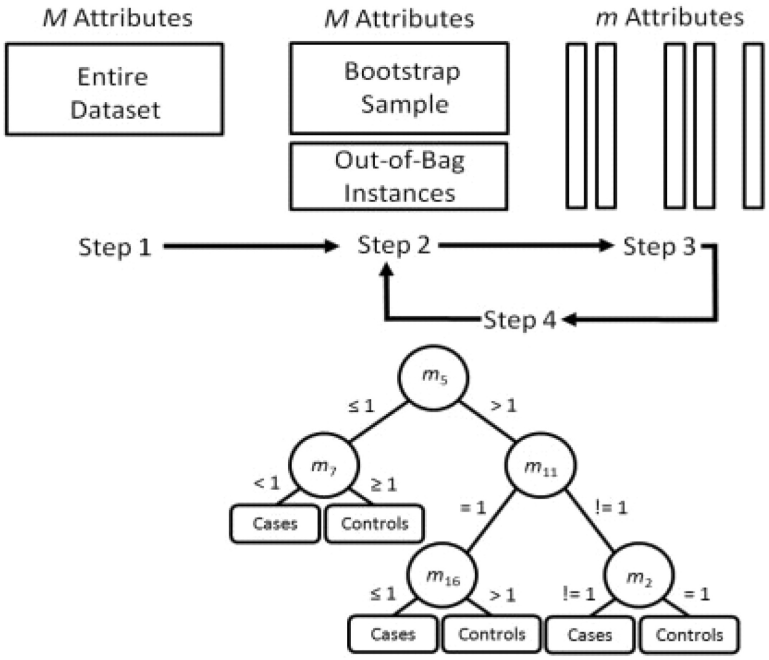

Figure 2.7: Iteration process to develop an ensemble of Decision Trees such as Random Forests or Extremely Randomized Trees classifiers. Image from "Bioinformatics challenges for genome-wide association studies" [4].

Neural Networks, more specifically, Deep Learning, is a technique based on the architecture of the biological brain, that turned out to be very good at discovering non-linear patterns in high-dimensional raw data, without much human supervision [79]. The unit of a Neural Network is the neuron, which is inter-linked with other neurons in multiple layers. There is always an input layer, which feeds the hidden layers, and that ultimately leads to the output layer. Deep Learning is considered to be a Neural Network with several hidden layers (more than 3). Each neuron's output is given by the weighted sum of outputs in the layers below, to which is applied a non-linear activation function. For example, the output of the j^{th} neuron

is given by:

$$f(a_{\mathrm{j}}) = f(\sum_{i} W_{\mathrm{ij}} X_{\mathrm{i}} + b_{\mathrm{i}}) \tag{2.13}$$

, where W is the weight of the neuron X and b is bias [80]. To perform backward propagation, the outputs are compared with the correct answer, and error derivatives are obtained. These are used to adjust the weights and improve the outputs of the network [80, 81]. Deep Neural Networks have produced extremely good results in the fields of image and language processing, and speech recognition. These challenges are similar in their high dimensionality and noise rates when compared to genotype-phenotype association problems. Some studies have used such methods to detect SNP interactions and perform GWAS in T2D, but they are very recent, albeit showing promising results [82, 83, 84].

DNN's have many benefits, but they also come with some drawbacks. For now, they lack an adequate formulation, and are considered black boxes, which makes it hard to interpret them. When uncovering associations in GWAS it is important to not lose track of the biological context of the problem, and that can be very hard when dealing with DNN's [79].

Overall, work flow of the processes used is roughly the same as shown in figure 2.8, and this is one of the aspects that this work aspires to change.

2.8 Validation of Machine Learning Methods

Machine Learning methods bring many advantages to GWAS, but they require proper validation so that we can be somewhat confident of their results. To perform validation, usually the available dataset is divided into a training set for the classifier to learn, and a test set, where it is analysed how it performs when introduced to new cases.

One method that is widely used to investigate overfitting is cross-validation. What this approach does is divide the dataset in k folds, and use them to train and test by turns. By using it, it is possible to verify how the validation metrics come up with different combinations of training and testing. If the classifier is indeed overfitting, the metrics will fluctuate a lot more, and some runs will output very poor results [5]. A visualization for better understanding of how this method operates can be seen on figure 2.9.

There are many metrics used to draw conclusions from these tests. Intuitively, one of

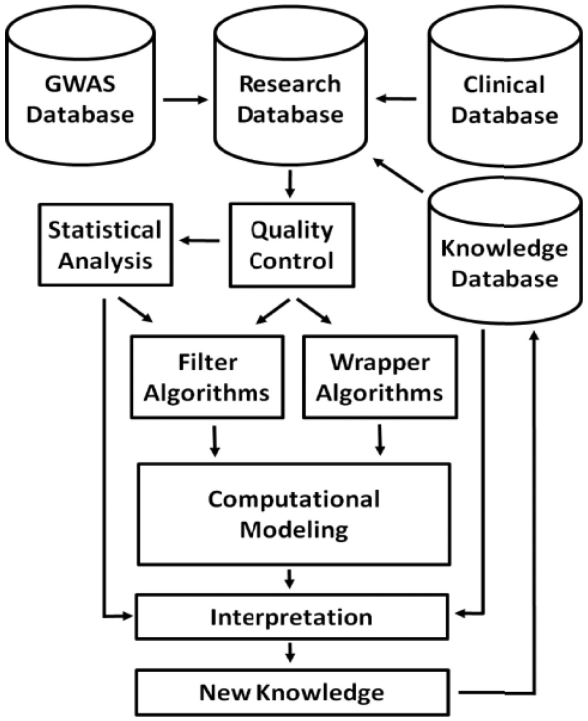

Figure 2.8: Regular work flow of projects that utilize computational methods in GWAS. Image from "Bioinformatics challenges for genome-wide association studies" [4].

the first metrics to look at is the percentage of predictions that were correctly made, designated accuracy. However, in some cases, it can be misleading and display values considered good, and the classifier stills turns out to be a poor one. This can happen for example, when the dataset is greatly unbalanced, and classifies only one of the classes correctly. To avoid this, it is also important to look at sensitivity (also called True Positive Rate or Recall) and specificity (or True Negative Rate). These metrics allow to understand how well the predictions went on both classes. Another one of these metrics used is the F1-Score (or F-measure), that combines both precision and recall to output an overall goodness-of-classification metric, although it fails to consider True Negatives [85]. All the previously described metrics calculations can be observed in figure 2.10.

The last metric covered, that is also very widely used to analyse the best decision thresholds are the Receiver Operating Characteristics. These are meant to evaluate the discriminant power of binary classifiers at different decision thresholds. For each threshold, the corresponding True Positive rate and False Positive Rate are plotted, which serves to find the best possible trade-off between these two metrics. Lastly, the Area Under Curve (AUC) is measured to get a picture of the global decision power for the predictor [6].

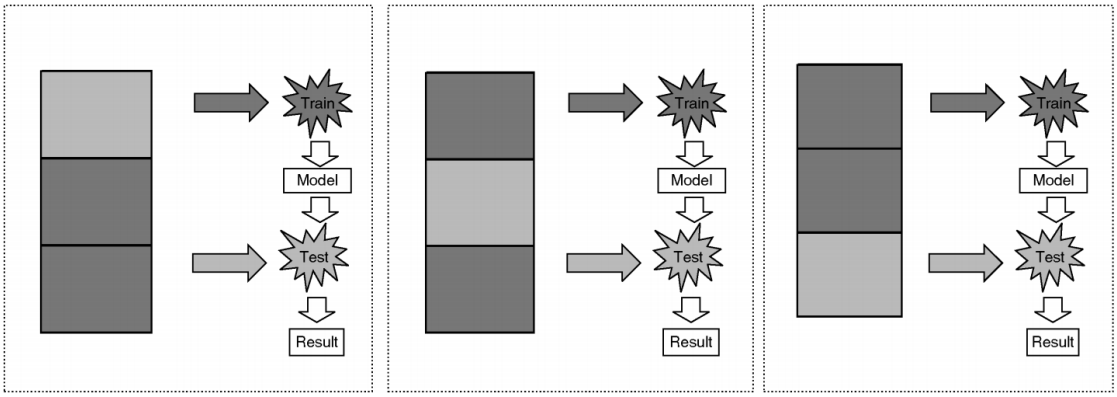

Figure 2.9: Demonstration of three-fold cross-validation. The same principle is applied to any k-fold cross-validation. Adapted from "Cross-Validation" [5].

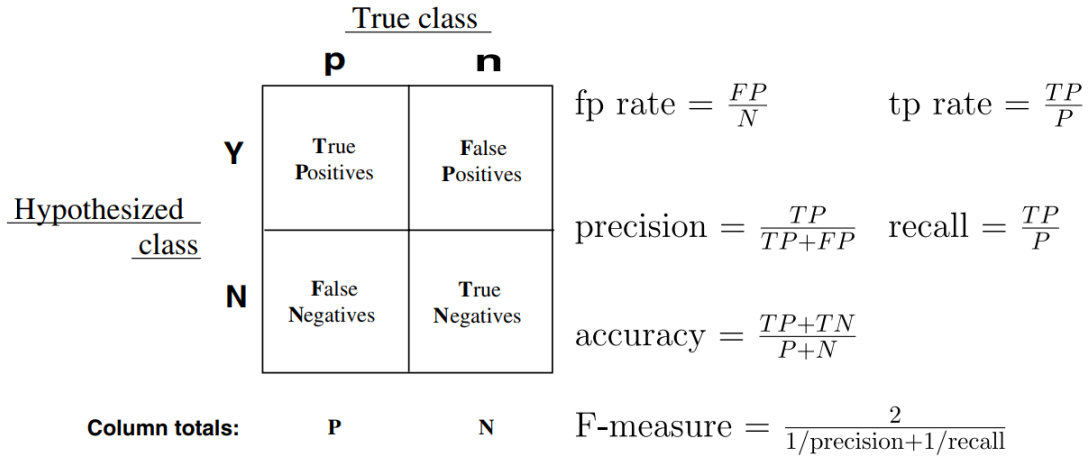

Figure 2.10: Several Machine Learning validation metrics and how to perform their calculations. Adapted from "An Introduction to ROC analysis" [6].

3

Data Preparation

3.1 Cases and Controls

To perform the association study, data from patients with T2D was collected. Both this data and the controls data is represented by VCF files. The cases VCF file contained 71 samples of the Portuguese patient's exome. When all of them were merged in a single file, there were 267 475 variants, either SNPs or INDELs. No data of patient medical records was used, since it is already known that features such as BMI and age can be great predictors of diabetes. Adding these to any model severely improves its accuracy, but the point of this study is to only use genomic data, more specifically SNPs, to both predict and find new markers for T2D.

Since in the cases file there was no information about sex, it was decided that the study would be conducted without this division. There are several diseases that have different risk metrics depending on sex, but for T2D, these are mostly due to environmental factors. As such, and also to not reduce further the number of samples by dividing it, division by sex was not performed.

The control data was collected from the 1000 Genome Project. It's goals involve discovering the most possible structural variants in the human genome of most ethnicities with frequencies of at least 1 %. This project was able to collect 2504 samples from 26 populations, with at least 4x genome coverage, genotyped with high accuracy [48]. One of those populations is the IBS, which is short for Iberian populations in Spain. Considering all the case samples are of Portuguese patients, the closest ethnic group possible was selected, to avoid bias in data that would eventually lead to differentiation of populations instead of T2D. The number of samples gathered from this study were 107. The data is divided by chromosome, with gzip compressed files ranging from 200 MB to 2 GB, totalling 17.4 GB of

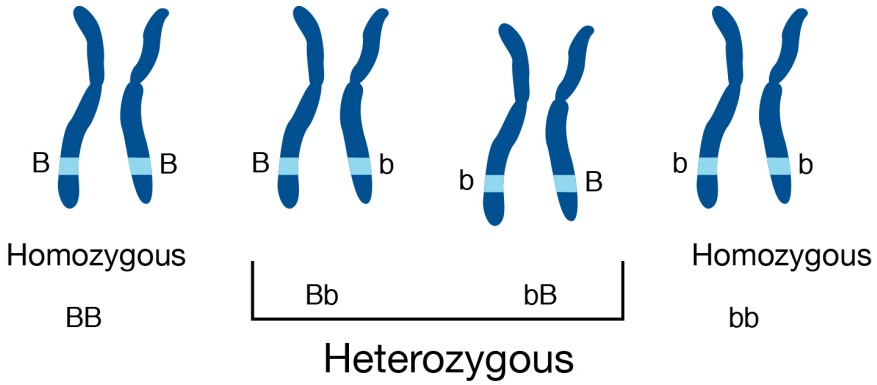

Figure 3.1: Visualization of possible genotypes for the structural variants. Adapted from public domain images at *https* : *//www.genome.gov* .

compressed data. When uncompressed, these files reach at least 500 GB of storage, containing over 80 million variants across the genome.

The cases VCF file offers numerous information of the variants, such as allele count in genotypes, total number of alleles in called genotypes, allele frequencies and so on. But the most important and the one that is going to be the most used is the GT field, which indicates the genotypes themselves. When making a call, short reads translate the information of two chromosomes. By using figure 3.1 as reference, and considering B as the REF allele, if most of the reads show a "B", it is genotyped as Homozygous by REF. If most show "b", it is considered homozygous by ALT. Since in this case, the genotype is unphased, it is not possible to know which chromosome the allele refers to, so both "Bb" and "bB" are classified as heterozygous. The REF allele is the one with the highest allele frequency in the population.

In VCF files, genotypes are represented as 0/0 for homozygous by REF, 0/1 for heterozygous, and 1/1 for homozygous by ALT. This needed to be translated in a way Machine Learning algorithms could interpret. To do so, a final structure of samples by variants is attained, where 0/0's are represented by 0's, 0/1's by 1's and 1/1's by 2's. The translation for SNPs with more than 1 ALT allele can be seen in the table 3.1.

3.2 Cases Quality Control

The controls data from the 1000 Genome Project is already guaranteed to be of very high quality. However, it is important to analyse the cases dataset quality to ensure the confidence level that is put into it. If we're not confident of the quality

GT	Translation
0/0	0
0/1	1
1/1	2
0/2	3
1/2	4
2/2	5
0/3	6
...	...

Table 3.1: Translation of the genotypes (GT) of the VCF file to standard samples by variant dataset, for SNPs with several ALT alleles.

of the dataset, several measures have to be taken to ensure results are not biased by wrong genotypes.

To analyse the quality of the cases dataset, a tool developed by Illumina called hap.py (*https* : *//github.com/Illumina/hap.py*) was used. This tool produces solid comparison metrics between two VCF files or one file and a reference genome. Since different genotyping methods can produce different ways of representing the structural variants, this comparison is not as straight forward as one might think. Hap.py besides solving these issues, counts SV types and produces quality metrics for them. However, when comparing files, it was noted that the tool only compared the first exome to the reference. This happens because it was built not considering multi-sample VCF files. As such, genotype accuracy metrics are not complete, but others such heterozygous to homozygous and transition to transversion ratios are still informative. The het/hom ratio is usually 2:1 for WGS and lower for WES. In the TsTv ratio, transitions are interchanges from purines or pyrimidines, and transversions involve interchanges of purines to pyrimidines or vice-versa. The expected proportion for this ratio is 2.1 for WGS and higher (up to over 3) for WES. The metrics for the first exome can then be observed in table 3.2. These were performed using the Platinum Genomes as the Gold Standard VCF file.

Type	Total Count	Truth het/hom	het/hom	Truth TsTv	TsTv
INDEL	3449	-	-	1.22	3.68
SNP	49073	1.51	1.68	2.09	2.44

Table 3.2: Total count of INDELs found on the Gold Standard Platinum Genomes, and truth and cases dataset het/hom and TsTv ratios. The counts are much lower than the total variants because these are the ones found in the truth set.

The het/hom ratio is slightly lower than 2:1 and the TsTv ratio is higher than the golden standard both because the dataset is of WES, which is the expected.

3.3 Dataset Construction

As of this point, there are two separate datasets with very distinctive number of variants. As such, it is necessary to merge them in a single file, guaranteeing the most possible number of variants in the final file.

This process is started by assembling the existing variants on the cases files and looking them up on the enormous chromosome files of the controls data. After all the possible variants are identified, their genotypes are extracted in the exact same way it was performed for the cases data. When lining up variants from both datasets, it was verified if their REF and ALT alleles matched. Those who did not were discarded. After it, 181 691 common variants were found between the two sets, which allowed to assemble them.

To finalize the data clean up, missing values needed to be handled. The cases file had a rate of 22% missing genotypes. This was improved when cases and controls were combined, but it was still a pressing issue. To solve it, every feature containing more than 10% missing genotypes was removed. The remaining ones with only a few data points missing were imputed utilizing the most frequent value in that column. This leads to a final total of 168 432 variants. A final "labels" column was added with 0's representing control samples and 1's representing the cases.

It is important to note that, although ethnicities are the most similar possible, cases and controls were most likely acquired by different sequencers which might introduce bias differentiating them. The number of samples and variants can be seen in the table 3.3.

Dataset	Number of Variants	Number of samples
cases	267 475	71
controls	81 271 745	107
total	181 691	178
imputed	168 432	178

Table 3.3: Number of samples and variants at each stage of the data processing.

3.4 From Variants to Genes

Although the data quality holds up, it is also necessary to assume that the dataset might contain some noise. This means that some variants can be extremely good differentiating the classes we created, especially since there many more variants than samples. These might not necessarily be noise or wrongfully genotyped variants, but it is better to look at whole regions and combine them instead, since it reduces this risk.

The reason why it is possible to infer information of variants from gene regions is thanks to Linkage Disequilibrium. As it can be seen on figure 3.2, by identifying regions of high LD it is possible to attain information of several disease related SNPs, even if they are not directly contained in the dataset.

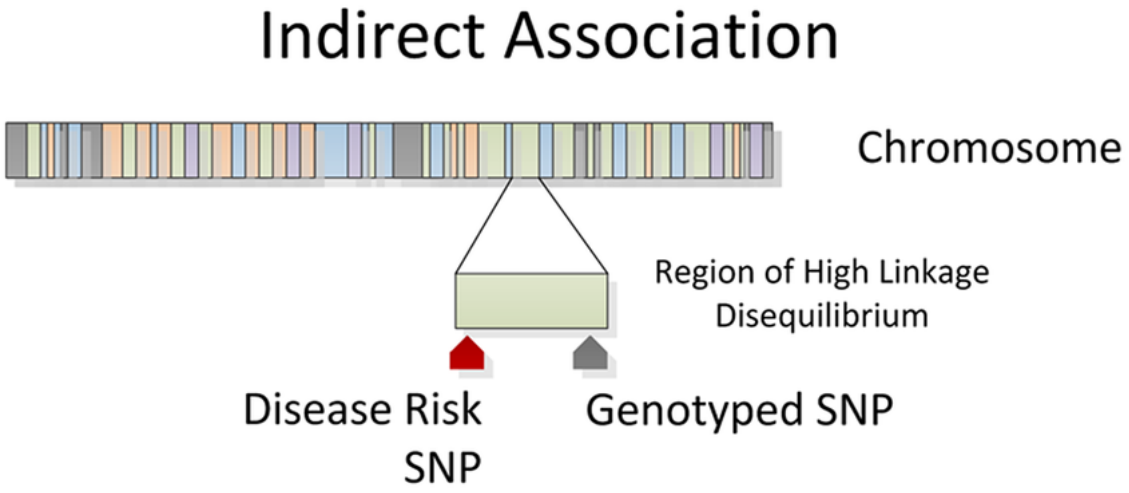

Figure 3.2: Visualization of high Linkage Disequilibrium regions which allow for usage of genotyped SNPs to infer disease risk SNP [7].

As such, it is critical to be able to extract information relative to which gene a variant belongs to, so it is possible to group them. This is one of the fastest and most meaningful ways of grouping variants, since directly performing tests of LD between each variant would be extremely computationally heavy.

To do so, a package for R called BiomaRt was used, that allows to query the Ensembl database. It works essentially as a genome browser with information of genomics, evolution, transcriptional regulation, sequence variation and annotations of genes, alignments and disease data. By querying the respective genes for the available variants, it is then possible to build a dictionary for an easy mapping of variants to genes. However, sometimes this querying cannot retrieve gene information because

the variant is still not mapped. When this happened, a note is attached to the variant, but the analysis progresses with these marked as having no gene.

4

Feature Engineering

4.1 Genes Pre-Selection

At this stage, there are 168 432 variants present in the dataset. One of the first possible approaches can be to perform feature reduction and follow up with classifiers training. However, with this extreme number of features and comparatively low samples, the likelihood of finding dataset specific patterns that are not scalable to alternative datasets is very high. As such, it is important to find a methodology that is able to adapt to different genotyping data and learn from either more features or samples.

As part of the methodology, and to standardize the approach, a translation of the current datasets from variants to genes was performed. This essentially leaves the data exactly as is, but provides extra information relative to the gene of each variant. This may come across as very little extra information added, but it allows to gather variants in groups, and distinguish with any extracted metric which are the meaningful variants, and the misfits in their group. Nonetheless, the main problem still remains. It is necessary to select which are the important genes.

To do so, two distinctive methods of feature selection are employed. The first one, makes use of prior knowledge of T2D genetics, and the second employs regular GWAS metrics combined with group information to establish if the genes are relevant or not.

A list of the most relevant genes discovered was obtained from "Genetics of Type 2 Diabetes" [1]. This is not a comprehensive list nor it has all of the variants discovered up to date, but it delivers important information such as the frequency of the risk allele in a population and the Odds Ratio for T2D, where it can be seen on table 4.1.

By looking at the OR of the genes in the table, it is possible to identify, as previously

Locus	Chr	Risk allele frequency	OR (95%CI)
NOTCH2	1	0.11	1.13 (1.08-1.17)
PROX1	1	0.5	1.07 (1.05-1.09)
IRS1	2	0.61	1.19 (1.13-1.25)
THADA	2	0.92	1.15 (1.10-1.20)
RBMS1/ITGB6	2	0.57	1.11 (1.08-1.16)
BCL11A	2	0.46	1.08 (1.06-1.10)
GCKR	2	0.62	1.06 (1.04-1.08)
IGF2BP2	3	0.29	1.17 (1.10-1.25)
PPARG	3	0.92	1.14 (1.08-1.20)
ADCY5	3	0.78	1.12 (1.09-1.15)
ADAMTS9	3	0.81	1.09 (1.06-1.12)
WFS1	4	0.27	1.13 (1.07-1.18)
ZBED3	5	0.26	1.08 (1.06-1.11)
CDKAL1	6	0.31	1.12 (1.08-1.16)
JAZF1	7	0.52	1.10 (1.07-1.13)
GCK	7	0.2	1.07 (1.05-1.10)
KLF14	7	0.55	1.07 (1.05-1.10)
DGKB/TMEM195	7	0.47	1.06 (1.04-1.08)
SLC30A8	8	0.75	1.12 (1.07-1.16)
TP53INP1	8	0.48	1.06 (1.04-1.09)
CDKN2A/B	9	0.79	1.20 (1.14-1.25)
TLE4	9	0.93	1.11 (1.07-1.15)
TCF7L2	10	0.25	1.37 (1.28-1.47)
HHEX	10	0.56	1.13 (1.08-1.17)
CDC123/CAMK1D	10	0.23	1.11 (1.07-1.14)
KCNQ1	11	0.61	1.40 (1.34-1.47)
KCNJ11/ABCC8	11	0.5	1.15 (1.09-1.21)
CENTD2	11	0.88	1.14 (1.11-1.17)
MTNR1B	11	0.3	1.09 (1.06-1.12)
HMGA2	12	0.1	1.10 (1.07-1.14)
TSPAN8/LGR5	12	0.23	1.09 (1.06-1.12)
OASL/HNF1A	12	0.85	1.07 (1.05-1.10)
PRC1	15	0.22	1.07 (1.05-1.09)
ZFAND6	15	0.56	1.06 (1.04-1.08)
FTO	16	0.45	1.15 (1.09-1.22)
HNF1B	17	0.43	1.12 (1.07-1.18)
DUSP9	X	0.12	1.27 (1.18-1.37)

Table 4.1: Collection of chromosome, risk allele frequency and OR of 37 risk genes identified for T2D. This list was adapted from "Genetics of Type 2 Diabetes" [1].

stated, that all these genes have relatively small values. Besides, these can only explain about 10% of the observed heritability of T2D [1]. The most prominent genes are the KCNQ1 and TCF7L2, them being the only ones with OR higher than 1.3. The OR is a metric indicative of association between a variable and an outcome of interest. Given the two-by-two frequency table 4.2 the OR is calculated through:

$$OR = \frac{a \times d}{b \times c} \tag{4.1}$$

	Diseased	Healthy
exposed	a	b
not-exposed	c	d

Table 4.2: OR two-by-two frequency table.

The closer OR is to one, the more indicative it is that the exposure does not affect the odds of the outcome. If it is higher, the exposure is associated with high odds for the outcome, and vice-versa. As such, and as it can be seen by the known T2D risk genes, these are not linked with moderate risks [86]. However, by using them together in a classifier, some non-linear relationships might be picked up, that are not clear with only OR.

Since the association between genes and variants was previously performed, the process of identifying the risk ones is very straightforward. However, since there was only data from WES, there are some risk genes that were not picked up in the current dataset. Out of the 37 genes, 18 were present, including some with higher OR such as KCNQ1 and DUSP9, but other also important ones like TCF7L2 and CDKN2A/B weren't. However, since one of the goals of this study is to also try and find new markers, it is necessary to include other genes. To do so, the standard association *Pearson's* χ^2 test is going to be used. It is then possible to group it's results to find more meaningful regions or genes. Even more so, since this data is nominal, the *Pearson's* χ^2 test is one of the most adequate for this situation. The Fisher test could also be applied, but as it can be seen further down, one metric is sufficient.

Before testing the variables, it is first necessary to test if they follow a normal distribution. Since Shapiro-Wilk normality test is not deemed accurate for over 50 samples, D'Agostino's K-squared test was used to test it, since it is the one performed by the "normaltest" function of the statistics python module Scipy. The test was applied to every variable, and the $-log_{10}(p - value)$ was calculated and

Figure 4.1: List of risk genes which were attempted to be found in the dataset. Genes present in the data are red coloured, and the remainders are silver. The bigger the size of a word, the bigger it's known OR to T2D according to "The genetic architecture of type 2 diabetes" [8].

plotted against each variant's position in the genome, as it can be seen on figure 4.2. Since the $-log_{10}(0.05) = 1.3$, and most of the values hugely surpass it, we assume all variants follow a normal distribution.

The next step, is to actually apply the *Pearson's* χ^2 test. This was performed in a similar way to the normality test. After it, we can clearly see in the figure 4.3 that some regions show higher correlation to the classes of T2D and healthy. However, by only choosing variants directly applying this metric, noisy or incorrectly genotyped are still going to be picked up.

Normally, to solve this issue, a Bonferroni correction is applied. If multiple tests are performed, the probability of a rare occurrence increases, and with it the likelihood of incorrectly rejecting a null hypothesis. As such, the α probability that a null hypothesis is rejected changes according to:

$$pi \leq \frac{\alpha}{m} \tag{4.2}$$

, where m is the number of tests completed. Instead of using it, a more region centred approach was used, that flags every gene where the average of every variant's p-value is less that 0.05. From it, 120 genes were selected, which greatly reduces the previous gigantic amount of features. By adding up the 18 identified T2D risk genes, the dataset ends up with 138 genes.

4.2 Feature Extraction

At this point, the dimension of the dataset has been greatly reduced, but it still contains data of variants that can single-handedly over fit the Machine Learning

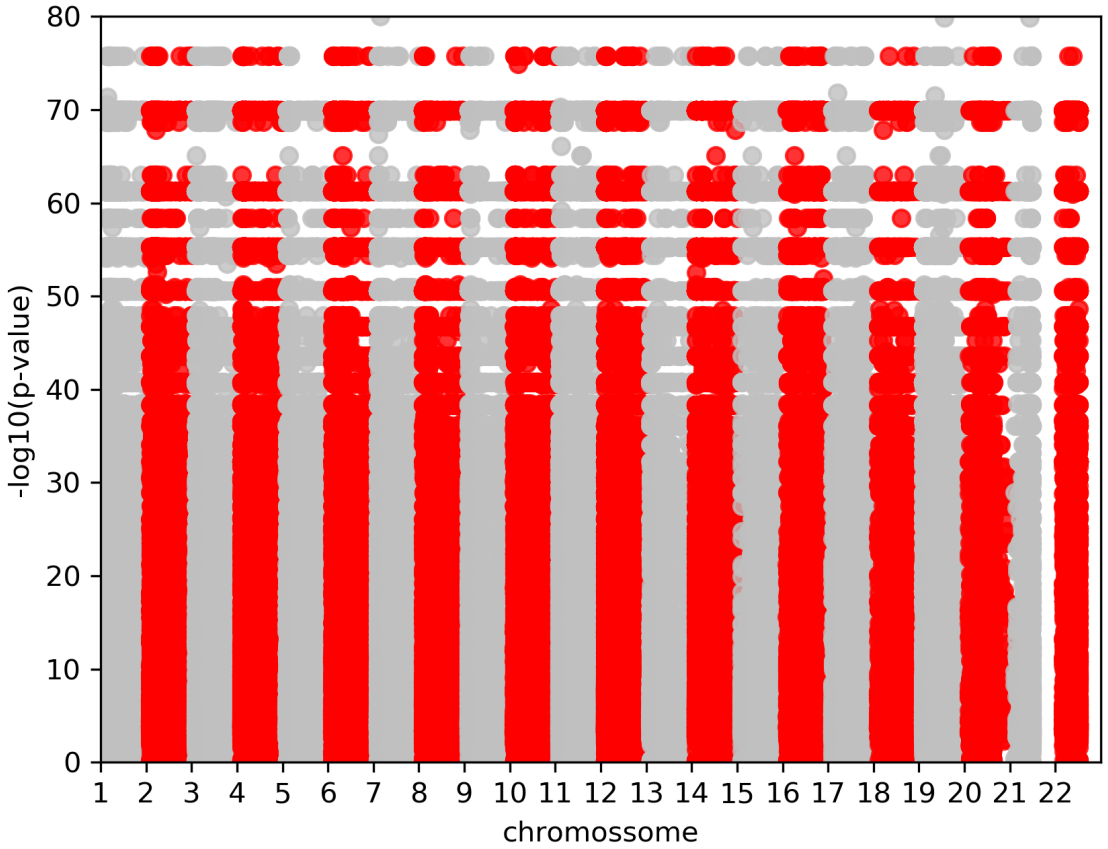

Figure 4.2: $-log10(p - value)$ of the D'Agostino's K-squared test for normality plotted against each variant's position in the chromosome.

models, especially since they are very likely to be overly correlated to both classes. So, rather than using them straight away, we can apply dimensionality reduction techniques, and by doing so, combine all the information and genotypes of one gene into a single dimension. Besides this, it is also possible to extract variance and mean of genotypes for each sample, to extract even more information.

The first dimensionality reduction method applied is the Principal Component Analysis. Since in this case, the projection only needs to be performed to one dimension, the only component that matters is the first one. For every single gene, this method is applied, which returns a vector of the first Principal Component, thus combining all the variants of each gene. The same strategy was applied with Linear Discriminant Analysis. This method assumes that the features are normally distributed, which was already tested beforehand. These two extracted features for each gene already capture most of the information contained by them, but to improve it even further, the mean and variance of genotypes by sample were also added.

Ultimately, the final dataset will be a combination of these 4 extracted features for

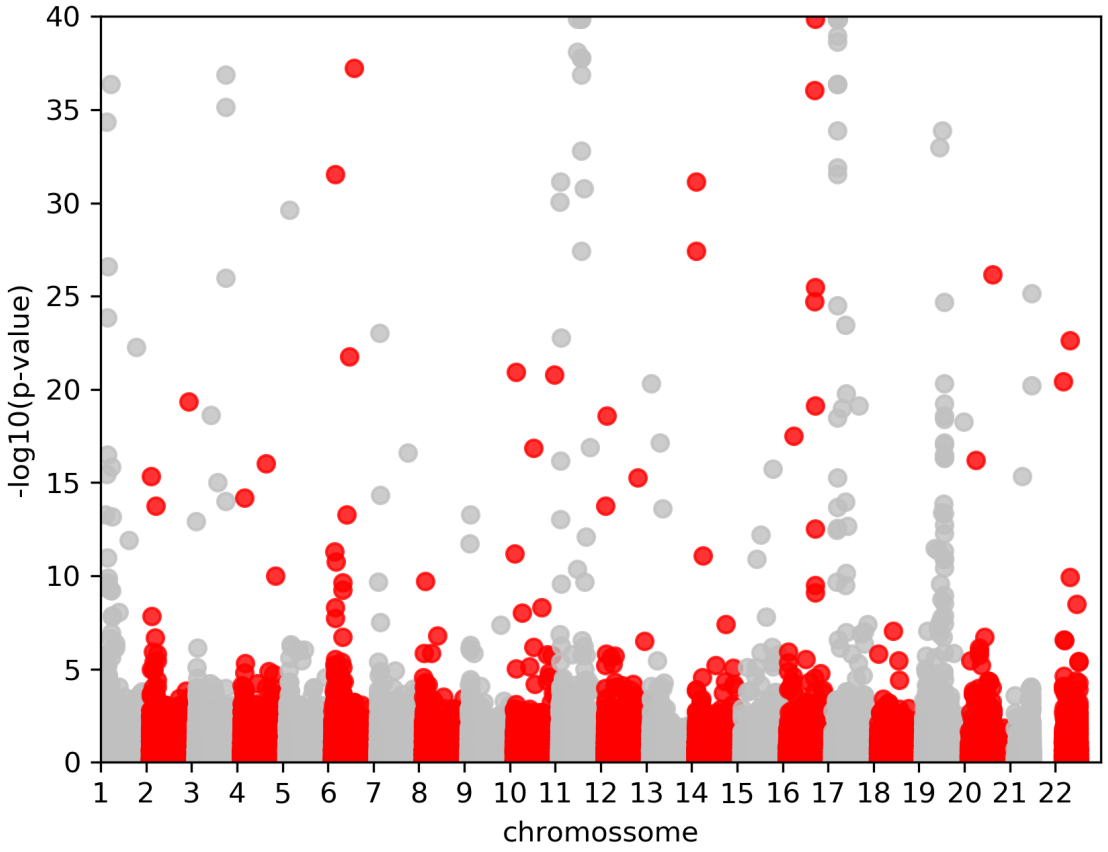

Figure 4.3: $-log10(p-value)$ of the *Pearson's* χ^2 test plotted against each variant's position in the chromosome.

each one of the originally selected 138 genes, which totals 552 features. These are named as "gene_method", so it is possible to identify the most relevant genes and feature extraction methods for each classifier.

4.3 Feature Reduction

To perform feature selection, we make use of an Extremely Randomized Trees ensemble (Extra-Trees for short), and it's ability to output which are the most important features. This classifier is an ensemble of Decision Trees, with the particularity of being much faster to train than Random Forests. These are among the most powerful Machine Learning algorithms available. The actual results and other metrics are not entirely relevant at this point, since only feature selection is being performed. This method can be seen in figure 4.5, which reflects the nodes of a simple decision tree applied to the final version of the dataset, with Gini impurity measure and the classes labelled as control for the healthy group, and target for the T2D affected

group.

To make use of this selection method, 1000 Extra-Trees classifiers were trained with the 552 features. For every classifier, the top 100 most important features were selected and counted. Using these counts, it is then possible to gather the frequency of which features are the most important when building an Extra-Trees classifier.

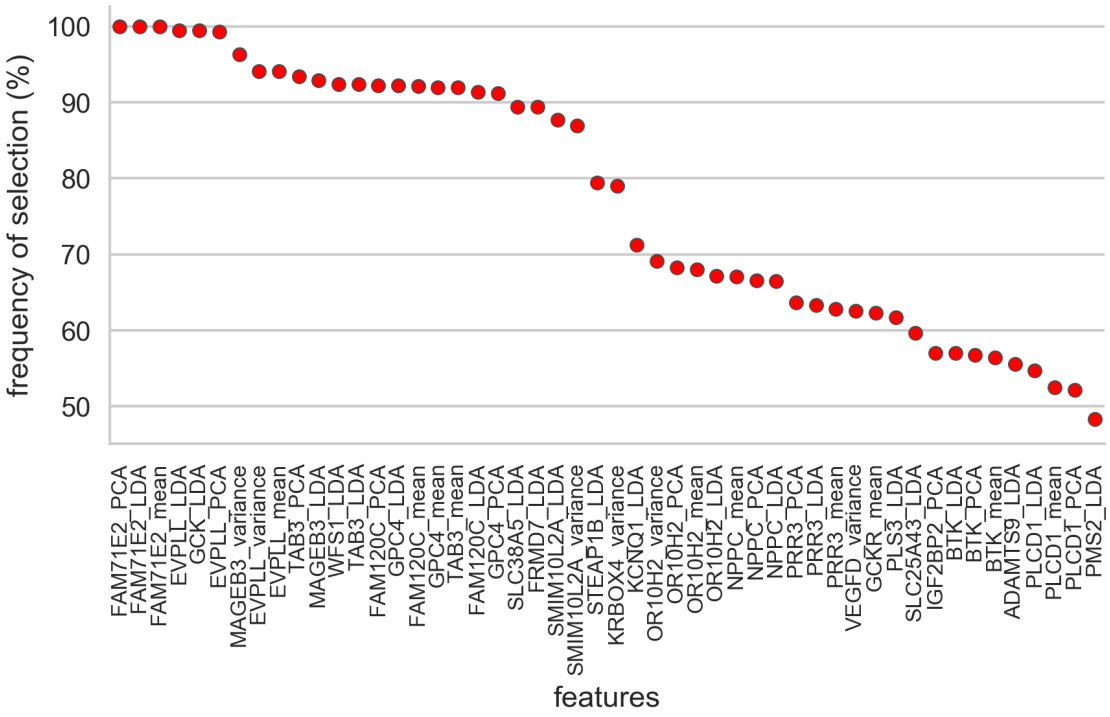

Figure 4.4: Frequency of times a feature was on the top 100 important features for each of the 1000 Extra-Trees classifiers trained. Top 50 features displayed.

From this test, features with over 50% frequency were selected, which amounts to a total of 49 features displayed on figure 4.4. Furthermore, from the selected features, we can verify that 25 genes are present, 6 of them already known for being linked to higher risk of T2D. Those risk genes are GCK, WFS1, KCNQ1, GCKR, IGF2BP2 and ADAMTS9.

Gene	Chromosome	Most Related Disease
FAM71E2	19	No results shown
EVPLL	17	Prostate Cancer
GCK	7	Maturity-Onset Diabetes of the young, Type 2 Diabetes Mellitus
MAGEB3	X	Melanoma
TAB3	X	Different Types of Cancer
WFS1		Wolfram Syndrome-1 Diabetes Mellitus
FAM120C	X	Autism
GPC4	X	Simpson-Golabi-Behmel Syndrome
SLC38A5	X	Pancreatic Ductal Adenocarcinoma
FRMD7	X	X-Linked Infantile Nystagmus
SMIM10L2A	X	No results shown
STEAP1B	7	Prostatitis
KRBOX4	X	Wilms Tumor 1
KCNQ1	11	Long Qt Syndrome 1 Diabetes Mellitus
OR10H2	19	No results shown
NPPC	2	Congestive Heart Failure
PRR3	6	No results shown
VEGFD	X	Different Types of Cancer
GCKR	2	Maturity-Onset Diabetes of the Young Diabetes Mellitus
PLS3	X	Osteoporosis
SLC25A43	X	Breast Cancer
IGF2BP2	3	Diabetes Mellitus, Noninsulin-Dependent
BTK	22	Agammaglobulinemia, X-Linked
ADAMTS9	3	Peters-plus syndrome Different Types of Cancer Body Mass Index Quantitative Trait Locus 11
PLCD1	3	Nail Disorder, Nonsyndromic Congenital, 3

Table 4.3: Ordered list of risk genes discovered and their most related diseases according to *www.malacards.org* , a human disease database.

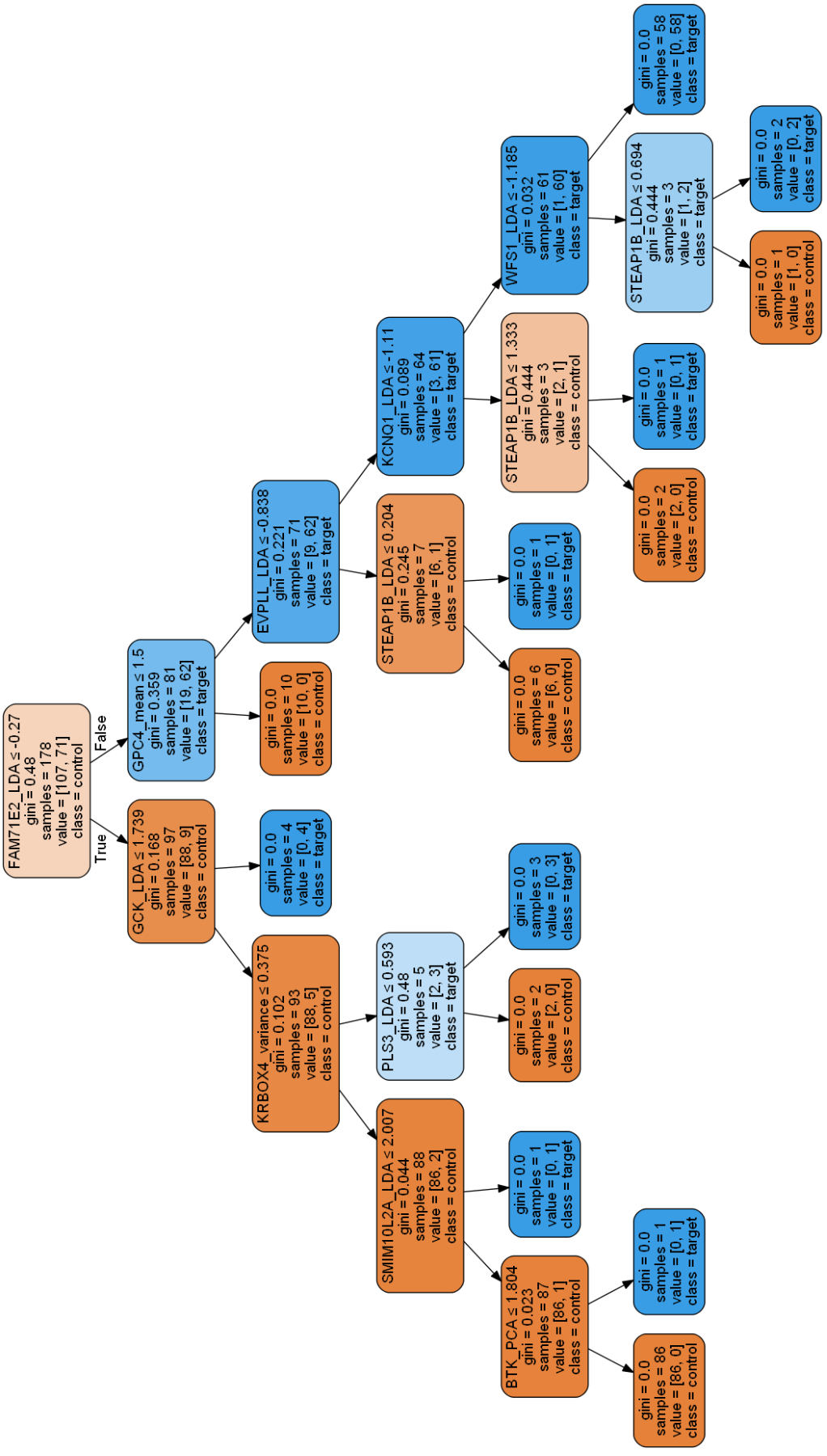

Figure 4.5: Decision Tree built with the dataset that contains all the features from the combined genes.

5

Classification and Results

5.1 Problem Formulation

Now that the dataset is entirely prepared, it has to be decided which classifiers are going to be used, what problem of classification are they actually tackling and how can the results be interpreted.

Firstly, it is important to understand what is it that it's going to be classified and how and if it can be extrapolated to the actual problem we are trying to solve. Our dataset is comprised of healthy people that make up the controls, and a group of T2D affected people, amounting the cases. Furthermore, these two groups were not necessarily genotyped by the same methods, or even sequenced by the same machines. As expected, this may lead to bias between the groups, which makes their classification easier but wrong. However, during the previous phases, certain aspects and mechanisms were already employed to deal or verify this situation. Foremost, by the analysis of the quality of the dataset, it is assumed that most of the genotypes of cases are correctly made and accordant with the reference genome, such as the controls groups knowingly is. Even if some variants are incorrectly called in either group, ultimately these are aggregated with many other variants in an attempt to represent a whole gene. This not only reduces the bias or noise the final features will have by dissolving the variants, but also allows for a representation that is understandable when those same features require further investigation.

One extra step employed to try and remove bias, was adding information of the biological context to the dataset. This was performed using the risk genes. By utilizing them, it is known beforehand that these do add up significant observable differences that are less likely to come from bias. When these were forcefully included in the dataset without any kind of filtering such as the one performed with χ^2, they only made up 18 out of 138 genes, or 13% of the genes. This number almost doubled

when there was a search for the most important features, which found 24% of the genes being already known as risk genes. This goes to show that the methods employed are indeed in the right track, and have a fair amount of success finding important and unbiased features.

Besides laying out the specific gene importance, that allows for follow up investigation on it, it is also possible to trace back the present variants on a single gene, and proceed with a closer inspection on specific SNPs and their genotypes. This nonetheless, still leads to the question of which is the correct approach for the problem, and what is it important to extract. Is it better to look at regions, or single variants are enough to explain the problem? Being able to detect mutations is important because it offers clear and concrete proof of what are the mechanisms behind the disease, and give credibility to the solution for genomics experts. However, this is not always possible, and if whole regions offer better results without specific alleles information, is it worthwhile to use them? For these questions, this approach enables answers for both, which might be important to justify using whole genes information not losing focus of smaller variants.

From the classification problem, it is expected to distinguish between groups of people with higher risk of T2D and healthy, only making use of genotypes. However, to specify the problem as such, there are a few assumptions that are required. First, that the genetic code does not undergo many changes during a person's lifetime [1] and that a member of the cases group has higher risk of developing T2D. The first assumption is acceptable to make, but there is no certainty on the second one. Since there is no access to physiological data of each patient, there is no telling which environmental conditions each is exposed to, therefore no specific information about habits versus heritability. Two people exposed to the same environmental conditions can have different outcomes regarding T2D because of genetics, but someone with lower risk can be affected by T2D solely through very high sugar levels in the blood. It is not known for sure if there are specific cases where genetics played a bigger role, but it is assumed that overall, the cases group has a higher risk of being affected by T2D. There is no specific interest in classifying T2D using more patient's data, but it would be helpful in such cases to perform a distinction between higher genetic T2D risk, and therefore improving the problem's approach.

Ultimately, the solution requires a classifier that is capable of looking for non-linear relations on the data. There are several available classifiers, but for this particular case Support Vector Machines and Decision Trees (more particularly Extra-Trees classifiers) were deemed appropriate. Both are well suited to look for non-linear

associations and handle medium-sized datasets. Deep Learning was also considered, but it is a black box where little information can be extracted about the biological context of the problem, and there are not enough samples to improve its performance in a meaningful way.

5.2 Classifier Optimization

There are many kernels and parameters that are possible for the classifiers, making it necessary to optimize them for the present dataset. For the optimization, a set number of parameters is given, and the best classifier is chosen based on the combinations of different parameters and a scoring function. Since this dataset is not terribly unbalanced, the scores are based on prediction accuracy. If it were, it would not be advisable to use accuracy as such, since it could be very high even if it misclassified one entire class.

For the SVM, the parameters provided are displayed on table 5.1. For each dataset, the optimized parameters will be different, which also happens when using only the top 50 discovered features, the whole dataset, or only features related to known risk genes. However, to standardize the process, only one final set of parameters that were the best overall for each dataset were chosen.

The C parameter, as explained previously refers to the number of violations of boundaries allowed, the tol is the tolerance for the stopping criterion, gamma is the coefficient for the kernels "poly", and "sigmoid and the degree refers to the polynomial degrees for the "poly" kernel. The final optimized SVM had a Gaussian Radial Basis Function for its kernel, a C of 0.75, tol of 0.001 and gamma set to "auto" that uses for it's value $\frac{1}{Number\ of\ features}$. The Gaussian RBF uses the following equation to compute the kernel:

$$K(a, b) = exp(-\gamma \|a - b\|^2) \tag{5.1}$$

, where a and b are the original vectors. The number of parameters tested is not more extensive since this would be extremely slow, even with the choice of combination of parameters being randomized.

The Extra Trees classifier parameters are displayed on table 5.2, and are different that the ones used before. The n_estimators refer to the number of trees in the forest and the criterion to the impurity evaluator. Other parameters refer to sin-

Parameters	Values
kernel	['linear', 'poly', 'rbf', 'sigmoid']
C	[0.25, 0.4, 0.5, 0.55, 0.75, 1]
tol	[1e-3, 1e-4, 1e-5]
gamma	[25, 50, 75, 100, 150, 'auto']
degree	[1, 2, 3, 5, 10]

Table 5.1: Parameters for the SVM classifier that were tested to find the most optimized one.

gle tree characteristics like min_samples_leaf to the minimum number of samples needed to form a node, min_samples_split to the minimum number to split a node and max_leaf_nodes to the maximum number of nodes on the tree (mostly to save memory).

The optimized predictor used 50 trees with the 'gini' criterion, at least one sample per node, a minimum of four to split a node and twenty max nodes per tree.

Parameters	Values
n_estimators	[25, 50, 75, 100]
criterion	['entropy', 'gini']
min_samples_leaf	[1, 2, 3, 5, 10]
min_samples_split	[2, 4, 5, 8, 10]
max_leaf_nodes	[2, 20, 50, 75, 100]

Table 5.2: Parameters for the Extra Trees classifier that were tested to find the most optimized one.

5.3 Metrics and Results

After optimization of classifiers, it is necessary to employ validation techniques and extract meaningful metrics that allow for higher confidence in the results of the classifiers. Nevertheless, these new cases are still from the same original dataset, and likely will have the same bias and noise. Although the ultimate metric to evaluate generalization is actual classification and validation in more datasets of the same type, this is not always possible.

For this particular case, 5 fold cross-validation was used and since in this problem the dataset is not terribly unbalanced, the tests were evaluated with the accuracy and F1-Scores. Since the F1-Scores don't account for True Negatives, the ROC curves

were also observed to select the best model possible. The Machine Learning python package Scikit-Learn combines unnecessary thresholds, which make some plots seem like they have fewer thresholds computed. By combining all these metrics, it is possible to get a good idea of how the classifiers are performing.

Since every metric, validation and classifier is decided, the plan for testing can now be formulated. There are three different datasets whose characteristics were previously detailed in the last chapter. These are the full dataset with all the features extracted for the selected genes, the dataset that only contains features related to known risk genes and the last one with the top 50 selected features. The original variants dataset was not used since it overfits very badly. To verify the consistency of results, for each dataset, a hundred classifiers were ran with five fold cross-validation (either SVM or Extra Trees), and their average values were computed. Their confidence intervals are not displayed since they are extremely small, because the predictors were very consistent. The metrics can be observed on figure 5.1 and the ROC curves on the following images.

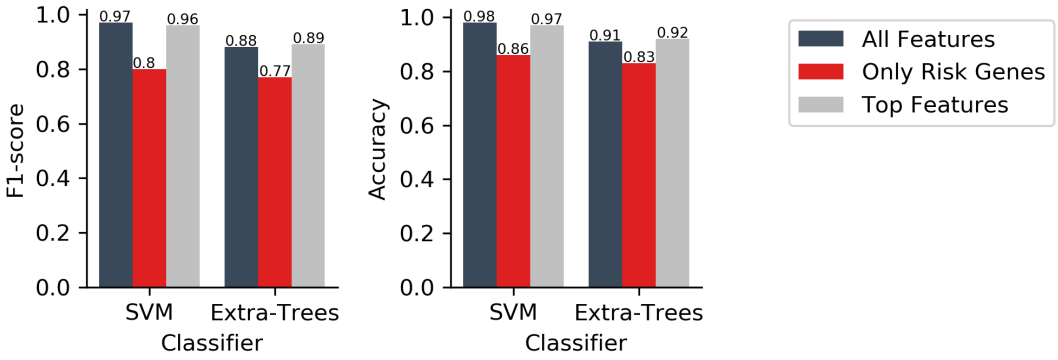

Figure 5.1: Mean of F1-Scores and accuracies by classifier built with all the features (blue), only features from known risk genes (red) and top features identified in the previous analysis (silver). Results are the average of 100 trained classifiers for each situation.

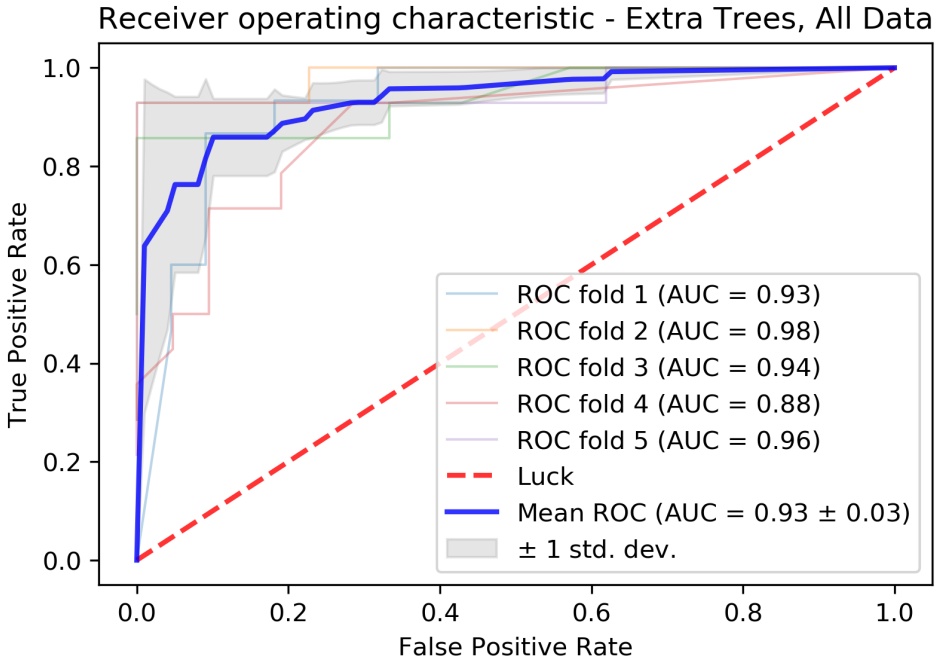

Figure 5.2: ROC curve for the dataset with all features, with the Extra Trees Classifier.

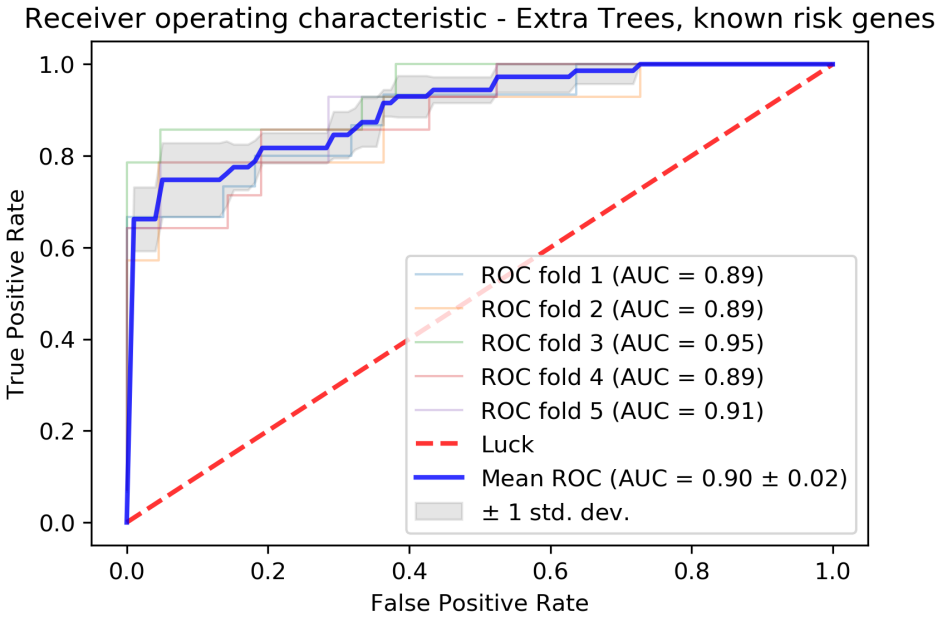

Figure 5.3: ROC curve for the dataset with the known risk genes features, with the Extra Trees Classifier.

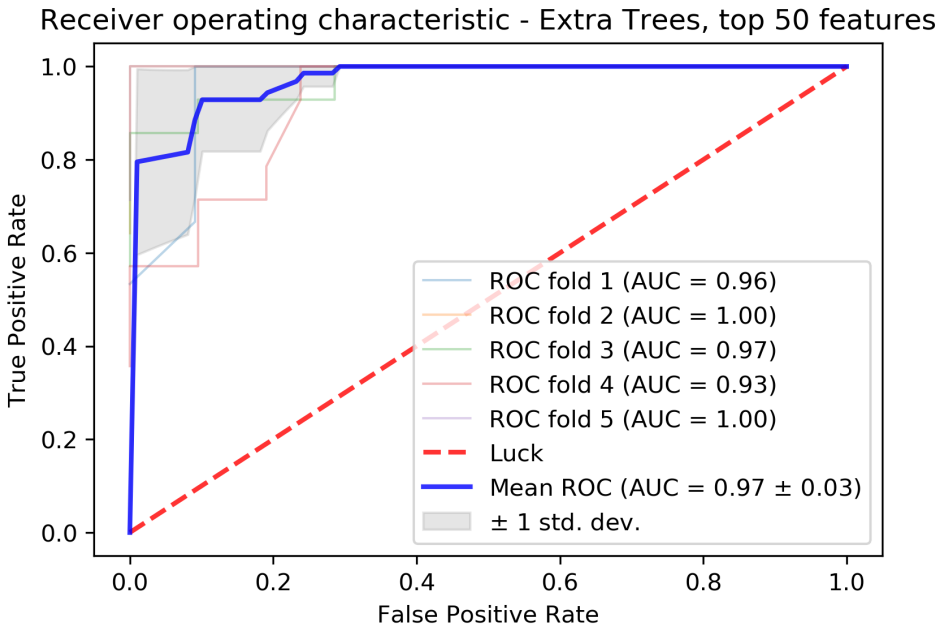

Figure 5.4: ROC curve for the dataset with the top fifty features, with the Extra Trees Classifier.

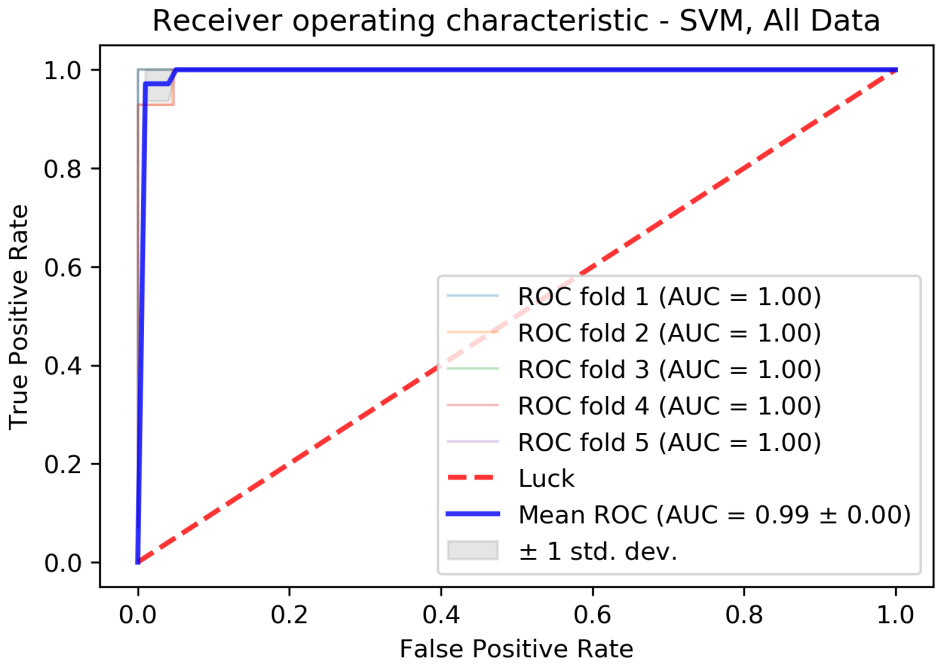

Figure 5.5: ROC curve for the dataset with all features, with the SVM Classifier.

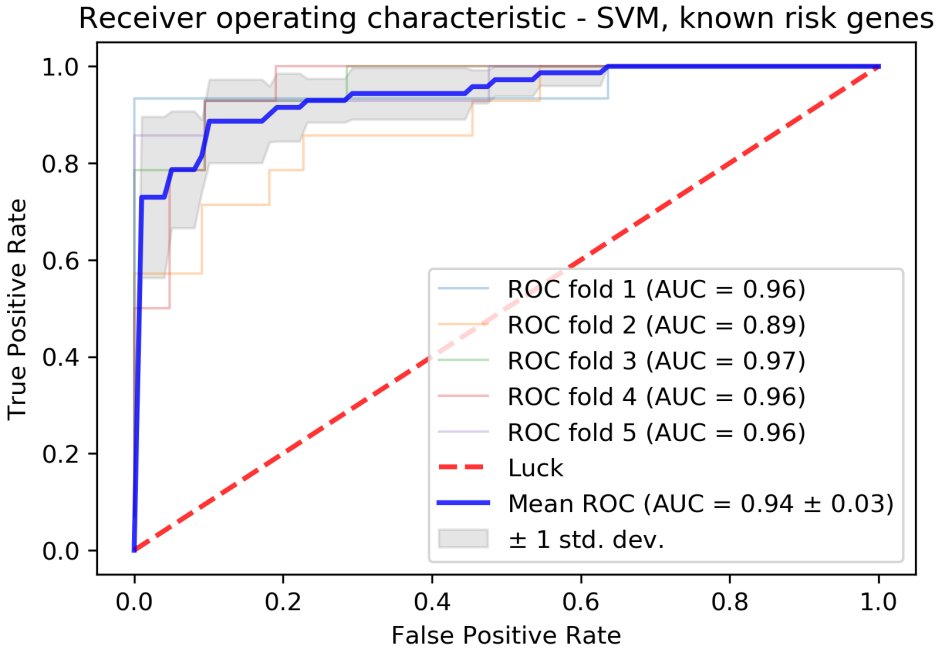

Figure 5.6: ROC curve for the dataset with the known risk genes features, with the SVM Classifier.

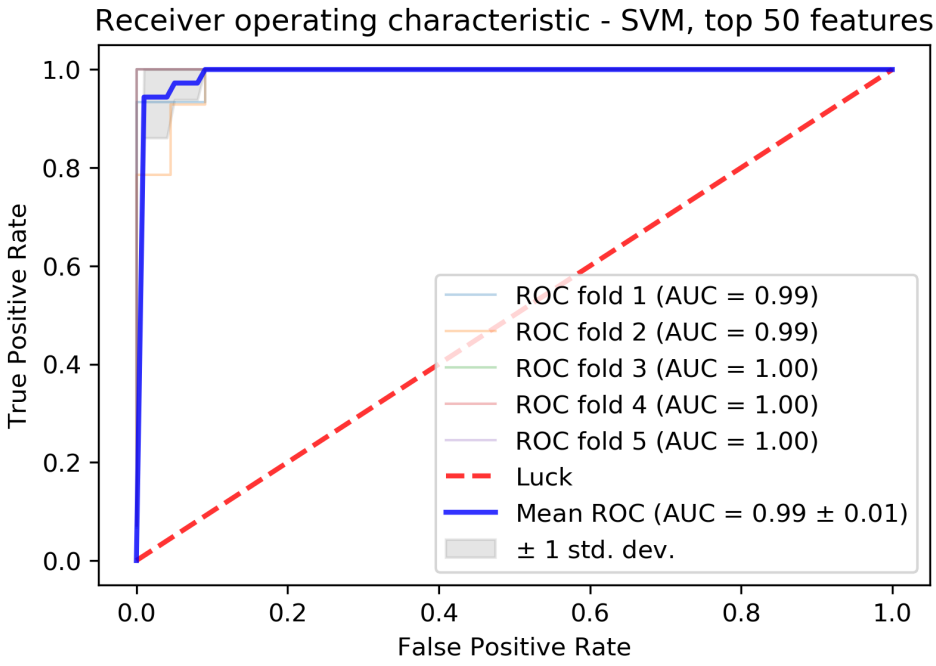

Figure 5.7: ROC curve for the dataset with the top fifty features, with the Extra Trees Classifier.

6

Discussion

6.1 Machine Learning in GWAS

The starting global premise of this work was fairly straightforward. It was intended from two discrepant datasets, to verify if it was possible to perform a T2D risk assessment. Given this, the single most important point in this project is the capacity to perform relevant feature extraction, and test non-linear relations between loci.

It is then very clear, that the main focus either from this or other GWAS is the feature engineering step. However, state of the art methods have some trouble with complex diseases such as T2D because and it is very easy to dismiss valid variants, since their correlation to the cases and controls labels are not always evident and is even non-linear [87]. As such, from the result of this work, it was possible to deliver a new feature engineering pipeline, that utilizes both filter and wrapper methods. This pipeline's results were then tested with machine learning methods that yielded extremely good results compared to any state of the art methods. The utility of this work is doubled when it was possible to identify some novel and interesting genes.

From the moment the genotype's dataset is prepared, the variants are combined into genes. Instead of performing filtering on SNP's, only the genes are used, which grants us the ability to look into whole regions. Performing gene-gene interactions for the whole dataset is not computationally feasible, but looking for regions with higher average correlation becomes now possible. By looking for genes with an average p-value of the standard χ^2 test below 0.05, some noisy features are reduced and it is ensured that whole regions are distinctive with high LD. After it, known risk genes from the literature can be added, which not only gives reliability, but also allows for easy modifications once the literature evolves. This sums up the first step of the feature engineering, that employs filtering methods and is reasonably standard, apart from looking at the problem from a region based perspective.

The second step of the pipeline, involves extracting new features from the existing selected genes. This allows for a targeted dimensionality reduction, providing only the relevant information from already relevant genes, well adapted for machine learning use. In this part of the process, specific gene information is mashed together in single highly detailed vectors, providing targeted depictions of the whole genes. At this point we are clearly stating the intent of using regions over single variants, as it enables to combine non-linear relations of the SNP's themselves, and allows further testing of non-linearity between genes and therefore, epistasis. This combination of factors is a great step forward from literature, as it now becomes possible to study relations that are very often overlooked since there are no good ways of measuring them. As an extra bonus point, all of these selections are specifically made and intended to be easily applied in Machine Learning models, that allow for straight away testing and validation.

To complete the pipeline, we arrive at the third and final feature selection segment. It's main goals are to maximize the prediction accuracy and find a relevant feature space that represents the problem with the least possible number of features. This section applies wrapper methods and decision trees to select those features, which is a process essentially based on Gini impurity and tree depth. From it, it is possible to identify several genes that can be novel introductions to T2D risk literature. In this step, it is also relevant to note that there was an increase of prevalence of known risk genes from the selected genes dataset (12.9%) to the top 50 identified genes (24%). We can then argue that the series of procedures utilized were indeed important to uncover T2D related risk genes.

The depiction of this process can be seen on figure 6.1, for an easier understanding of the work flow. From what it seems like a simple and straightforward pipeline, many new relations can now be studied, as it allows for a maximum information retention and selection without losing biological context.

After this process, the data was classified using SVMs and Extremely Randomized Trees classifiers. It is clear that the SVMs with the Gaussian Radial Basis Function kernel were the better performers in every dataset. Also, even though there was a reduction of features to less than 10% of the original size of the full dataset, only 1% of accuracy and f1-score were lost. Five-fold cross-validation was then used as a measure to avoid overfitting, and so that we can be more confident in the results. However, the point that provides the most confidence and validity over the pipeline is the 0.94 ± 0.03 AUC with a SVM and only known risk genes being used, that shows it can produce accurate classification based on known genes that increase the

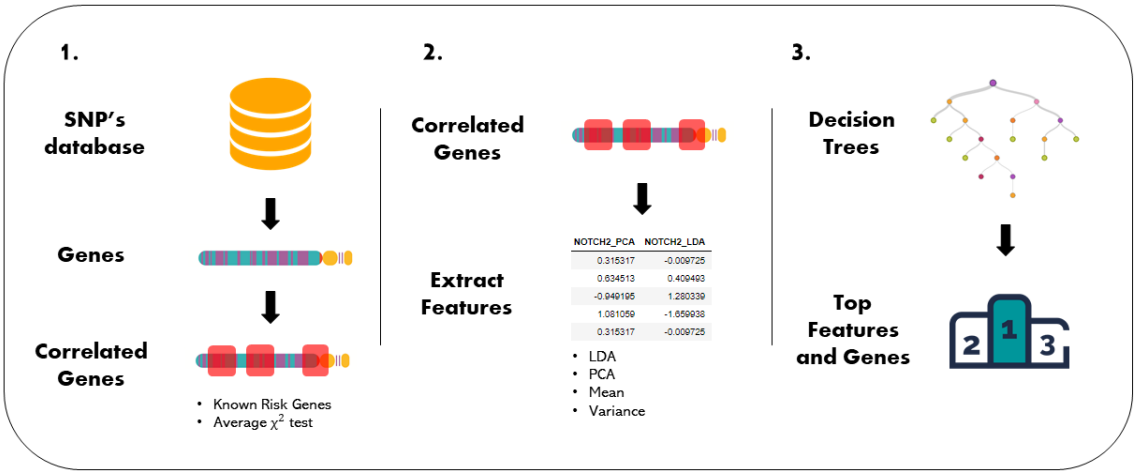

Figure 6.1: Depiction of the pipeline developed to extract important features and discover possible risk genes. On 1. the passage of variants/SNPs to finally relevant genes is shown. On 2. it is demonstrated the passage of those genes to features extracted and on 3. the ranking of variables with Decision Trees.

risk of T2D.

At the beginning, as stated before, with the variants and genotypes data, all the classifiers would massively overfit and classify every class correctly, mostly based on incorrect or noisy data. This pipeline, not only provides a way to employ Machine Learning in GWAS, but also to correct these issues, be more confident on the data used, and extract novel genes information.

The best SVM classifier can also act as risk predictor for this disease, and even output probabilities for a new sample of T2D risk. For this purpose, we assume 100 % to be the highest risk possible, but this is only based on the current data of what the model depicts as the higher risk possible. Nonetheless, these probabilities are only model based, and to develop a truly accurate risk predictor, it is first necessary to get a better understanding of the underlying genetics of T2D and make a more truthful representation of the phenotype differences of each patient.

6.2 Model Shortcomings

The results of the pipeline are fairly solid and after the dataset quality tests and validation metrics applied, we can be somewhat confident in them. However, there are a few points that the model fails to handle, and some pitfalls that can bias the results.

The first such point is that the sequencing machines utilized are not the same for the cases and controls groups. The same can be said about the genotyping methods utilized. Nonetheless, it is expected a certain level of confidence and a standardization of the procedures employed, so that the quality of the data is not affected, and the machine from which the genome or exome was sequenced does not matter. To even lessen the variability possibilities, samples from the same ethnicities were used, all to ensure the complete dataset was as reliable as possible. However, the use of samples from only the Iberian Peninsula can also be considered a drawback, as it makes the study plenty restricted in terms of world population reachability. Nevertheless, the pipeline still holds true for any other datasets it can be tested on.

The next shortcoming is relative to the ability of the pipeline to retrieve information of the variants and respective alleles that are responsible for the risk alterations that are seen. This retracing is not impossible, but it was discarded for the advantage of using regions. Since the genes information could be collapsed in single vectors, it was more useful for a machine learning model to access it rather than single variants. The retracing can be performed by looking at the intended original gene and performing association tests, but it is possible that some information is being lost or several variants are needed because of their non-linear interactions. Not providing risk allele information might very well be a pitfall, since many sources expect this kind of information.

Lastly, it is important to discuss the novel risk genes that were identified, their validity, and why are genes non-related to T2D considered by the top classifiers. On the first place, since there is no access to the patients data besides their genome, it is not possible to verify if there are any other conditions that might be affecting the results, and if they justify the appearance of many other genes related to risk of other conditions. Nonetheless, only the known risk genes and genes that have no further data available related to disease, already make up for 40% of the top 25 genes identified in the top 50 features. Even if for a moment we consider those non-related extra genes as noise, it doesn't take away from the fact that the classifiers fare up decently well with only known risk genes. The classifiers look for the best possible

data to classify the problem which may lead to usage of noise that fits the classes. To ascertain the validity of some possible novel identified genes, further gene tests would need to be performed.

So far, the results and validation metrics seem to weight up a positive outlook, but further certainty of this pipeline will not be entirely known until further validation can be performed with other SNPs or variants datasets.

7

Conclusion

7.1 Analysis Pipeline

The first problem that was dealt with in this project was the sheer size of the datasets, almost in the Terabyte order. This was a massive challenge, and required great optimization on the part of any code that handled them. Not only this, but they required very careful parsing as any bias introduced could ruin the results and the validity of any methods tested. Nevertheless, this task was successful, as proven by the quality control later performed on the dataset.

At the start of this work, it wasn't decided that a pipeline was going to be developed. The first main goal intended to establish a risk predictor for T2D and discover new genetic markers for it. However, as the project went on, it was observed that finding novel markers goes hand-in-hand with feature selection and that it really plays a big part when it comes to GWAS. It was also noted that for many classifiers in the literature, feature selection remained largely the same of any regular GWAS.

As such, I believe the pipeline of feature engineering that was developed to be of great use when exploring any new genomics dataset, with all the advantages and disadvantages discussed earlier. Besides, any new markers that are discovered with it can then be put to a test by classifying T2D, which gives more confidence on them.

The final risk predictor reached extremely good results, much better than expected or found in the literature. This was a bit of a surprise, and caused for many points of doubt or distrust in the process. However, at the end, most faults possible that may have happened were thought of, tested, and taken into account through all the steps, to finally achieve a healthy trust on the conclusions provided.

If this pipeline can be validated on other dataset, it will show that working with gene regions is a very important approach that needs to be incorporated when

performing complex diseases studies, and confirm that indeed epistasis might be the phenomenon that was missing to be accounted for to explain the missing heritability of complex diseases.

7.2 Future Work

The most important future work that can have the greatest impact, is the validation of this pipeline on bigger and more diverse datasets, Whole Genome data or even datasets of other diseases. Not only this, but also having more data of every patient can help forming groups or identify other points that can skew the labels. If this was possible, it could shift the way genetic datasets are interpreted, and accelerate the introduction of Artificial Intelligence use in these kinds of problems. To make this happen in the health field, the methods need to be transparent, understandable, and very clearly transmitted.

The following approach that could be linked to this study, are gene expression datasets. This would include perhaps a new whole analysis and data acquisition, but could add extra validity to the results of this pipeline.

Ultimately, novel methods that insist on the same points of region analysis, non-linearity and that consider epistasis can also be developed, because as it was shown with this one, they can provide a great deal information and select a good feature space to predict complex disease risk. These pipelines can then be integrated in an ensemble, to perform risk assessment and advance the use of such algorithms in personal health.

7.3 Personal Note

Since my first classes of programming in the first grade of Biomedical Engineering, I thought that I might have made a wrong decision in what comes to degree choice. However, since I was many times in contact with programming, and further along with data analysis and Machine Learning, I've in this way, and I'm so very glad that I did.

Currently, data scientists are in high demand, and very rightfully so, as their power to extract value from datasets is incredible. The same can be applied to the Life Sciences and Personal Healthcare areas, which to this day, are the ones that interest

me the most. With further work for validation and trust for these methods from the medical communities, there are many great areas where an impact can be made, such as this one.

Through and through, I really enjoyed putting my sweat into this endeavour, and it served to show the amazing things that are possible in the fields of Bioinformatics. It makes me very happy that in this line of work I am not bound by anything, except for a great deal of effort, and lots and lots of computing power.

Bibliography

[1] O. Ali, "Genetics of type 2 diabetes," *World journal of diabetes*, vol. 4, no. 4, p. 114, 2013.

[2] F. Sanger, S. Nicklen, and A. R. Coulson, "Dna sequencing with chain-terminating inhibitors," *Proceedings of the national academy of sciences*, vol. 74, no. 12, pp. 5463–5467, 1977.

[3] A. Géron, *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems.* " O'Reilly Media, Inc.", 2017.

[4] J. H. Moore, F. W. Asselbergs, and S. M. Williams, "Bioinformatics challenges for genome-wide association studies," *Bioinformatics*, vol. 26, no. 4, pp. 445–455, 2010.

[5] P. Refaeilzadeh, L. Tang, and H. Liu, *Cross-Validation*, pp. 532–538. Boston, MA: Springer US, 2009.

[6] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[7] W. S. Bush and J. H. Moore, "Genome-wide association studies," *PLoS computational biology*, vol. 8, no. 12, p. e1002822, 2012.

[8] C. Fuchsberger, J. Flannick, T. M. Teslovich, A. Mahajan, V. Agarwala, K. J. Gaulton, C. Ma, P. Fontanillas, L. Moutsianas, D. J. McCarthy, *et al.*, "The genetic architecture of type 2 diabetes," *Nature*, vol. 536, no. 7614, pp. 41–47, 2016.

[9] X. Zhang, S. Huang, Z. Zhang, and W. Wang, "Mining genome-wide genetic markers," *PLoS computational biology*, vol. 8, no. 12, p. e1002828, 2012.

[10] D. J. Witherspoon, S. Wooding, A. R. Rogers, E. E. Marchani, W. S. Watkins, M. A. Batzer, and L. B. Jorde, "Genetic similarities within and between human populations," *Genetics*, vol. 176, no. 1, pp. 351–359, 2007.

[11] I. H. . Consortium *et al.*, "Integrating common and rare genetic variation in diverse human populations," *Nature*, vol. 467, no. 7311, p. 52, 2010.

[12] S. Chatterjee, K. Khunti, and M. J. Davies, "Type 2 diabetes," *The Lancet*, vol. 389, no. 10085, pp. 2239–2251, 2017.

[13] R. B. Prasad and L. Groop, "Genetics of type 2 diabetes—pitfalls and possibilities," *Genes*, vol. 6, no. 1, pp. 87–123, 2015.

[14] G. Willemsen, K. J. Ward, C. G. Bell, K. Christensen, J. Bowden, C. Dalgård, J. R. Harris, J. Kaprio, R. Lyle, P. K. Magnusson, *et al.*, "The concordance and heritability of type 2 diabetes in 34,166 twin pairs from international twin registers: the discordant twin (discotwin) consortium," *Twin Research and Human Genetics*, vol. 18, no. 6, pp. 762–771, 2015.

[15] D. K. Sanghera and P. R. Blackett, "Type 2 diabetes genetics: beyond gwas," *Journal of diabetes & metabolism*, vol. 3, no. 198, 2012.

[16] X. Wang, G. Strizich, Y. Hu, T. Wang, R. C. Kaplan, and Q. Qi, "Genetic markers of type 2 diabetes: Progress in genome-wide association studies and clinical application for risk prediction," *Journal of diabetes*, vol. 8, no. 1, pp. 24–35, 2016.

[17] S. Vijan, "Type 2 diabetes," *Annals of internal medicine*, vol. 152, no. 5, pp. ITC3–1, 2010.

[18] J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, *et al.*, "Common snps explain a large proportion of the heritability for human height," *Nature genetics*, vol. 42, no. 7, p. 565, 2010.

[19] D. E. Reich and E. S. Lander, "On the allelic spectrum of human disease," *TRENDS in Genetics*, vol. 17, no. 9, pp. 502–510, 2001.

[20] O. Zuk, E. Hechter, S. R. Sunyaev, and E. S. Lander, "The mystery of missing heritability: Genetic interactions create phantom heritability," *Proceedings of the National Academy of Sciences*, vol. 109, no. 4, pp. 1193–1198, 2012.

[21] E. S. Lander, "Initial impact of the sequencing of the human genome," *Nature*, vol. 470, no. 7333, p. 187, 2011.

[22] J. R. Huyghe, A. U. Jackson, M. P. Fogarty, M. L. Buchkovich, A. Stančáková, H. M. Stringham, X. Sim, L. Yang, C. Fuchsberger, H. Cederberg, *et al.*, "Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion," *Nature genetics*, vol. 45, no. 2, p. 197, 2013.

[23] C. Bommer, E. Heesemann, V. Sagalova, J. Manne-Goehler, R. Atun, T. Bärnighausen, and S. Vollmer, "The global economic burden of diabetes in adults aged 20–79 years: a cost-of-illness study," *The lancet Diabetes & endocrinology*, vol. 5, no. 6, pp. 423–430, 2017.

[24] O. da Diabetes, "Diabetes: Factos e números-o ano de 2015-relatório anual do observatório nacional para a diabetes," *Lisboa: Sociedade Portuguesa de Diabetologia*, 2016.

[25] J. M. Chan, E. B. Rimm, G. A. Colditz, M. J. Stampfer, and W. C. Willett, "Obesity, fat distribution, and weight gain as risk factors for clinical diabetes in men," *Diabetes care*, vol. 17, no. 9, pp. 961–969, 1994.

[26] S. M. Haffner, "Epidemiology of type 2 diabetes: risk factors," *Diabetes care*, vol. 21, no. Supplement 3, pp. C3–C6, 1998.

[27] W. H. Organization *et al.*, *Global report on diabetes*. World Health Organization, 2016.

[28] S. H. Ley, O. Hamdy, V. Mohan, and F. B. Hu, "Prevention and management of type 2 diabetes: dietary components and nutritional strategies," *The Lancet*, vol. 383, no. 9933, pp. 1999–2007, 2014.

[29] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, *et al.*, "The sequence of the human genome," *science*, vol. 291, no. 5507, pp. 1304–1351, 2001.

[30] E. C. Hayden, "Is the $1,000 genome for real?," *Nature News*, 2014.

[31] J. MacArthur, E. Bowler, M. Cerezo, L. Gil, P. Hall, E. Hastings, H. Junkins, A. McMahon, A. Milano, J. Morales, *et al.*, "The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog)," *Nucleic acids research*, vol. 45, no. D1, pp. D896–D901, 2016.

[32] K. Läll, R. Mägi, A. Morris, A. Metspalu, and K. Fischer, "Personalized risk prediction for type 2 diabetes: The potential of genetic risk scores," *Genetics in Medicine*, vol. 19, no. 3, p. 322, 2017.

[33] W. J. Ansorge, "Next-generation dna sequencing techniques," *New biotechnology*, vol. 25, no. 4, pp. 195–203, 2009.

[34] M. Choi, U. I. Scholl, W. Ji, T. Liu, I. R. Tikhonova, P. Zumbo, A. Nayir, A. Bakkaloğlu, S. Özen, S. Sanjad, *et al.*, "Genetic diagnosis by whole exome capture and massively parallel dna sequencing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 19096–19101, 2009.

[35] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, *et al.*, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, p. 747, 2009.

[36] J. M. Zook, B. Chapman, J. Wang, D. Mittelman, O. Hofmann, W. Hide, and M. Salit, "Integrating human sequence data sets provides a resource of benchmark snp and indel genotype calls," *Nature biotechnology*, vol. 32, no. 3, p. 246, 2014.

[37] S. McCarthy, S. Das, W. Kretzschmar, O. Delaneau, A. R. Wood, A. Teumer, H. M. Kang, C. Fuchsberger, P. Danecek, K. Sharp, *et al.*, "A reference panel of 64,976 haplotypes for genotype imputation," *Nature genetics*, vol. 48, no. 10, p. 1279, 2016.

[38] L. J. Scott, K. L. Mohlke, L. L. Bonnycastle, C. J. Willer, Y. Li, W. L. Duren, M. R. Erdos, H. M. Stringham, P. S. Chines, A. U. Jackson, *et al.*, "A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants," *science*, vol. 316, no. 5829, pp. 1341–1345, 2007.

[39] M. Slatkin, "Linkage disequilibrium—understanding the evolutionary past and mapping the medical future," *Nature Reviews Genetics*, vol. 9, no. 6, pp. 477–485, 2008.

[40] K. Pearson, "X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 50, no. 302, pp. 157–175, 1900.

[41] C. S. Haley, S. A. Knott, *et al.*, "A simple regression method for mapping quantitative trait loci in line crosses using flanking markers," *Heredity*, vol. 69, no. 4, pp. 315–324, 1992.

[42] W. Hill and A. Robertson, "Linkage disequilibrium in finite populations," *TAG Theoretical and Applied Genetics*, vol. 38, no. 6, pp. 226–231, 1968.

[43] P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang, "10 years of gwas discovery: biology, function, and translation," *The American Journal of Human Genetics*, vol. 101, no. 1, pp. 5–22, 2017.

[44] M. Mooney, B. Wilmot, S. McWeeney, *et al.*, "The ga and the gwas: using genetic algorithms to search for multilocus associations," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 9, no. 3, pp. 899–910, 2012.

[45] S. D. Turner, R. L. Berg, J. G. Linneman, P. L. Peissig, D. C. Crawford, J. C. Denny, D. M. Roden, C. A. McCarty, M. D. Ritchie, and R. A. Wilke, "Knowledge-driven multi-locus analysis reveals gene-gene interactions influencing hdl cholesterol level in two independent emr-linked biobanks," *PloS one*, vol. 6, no. 5, p. e19586, 2011.

[46] C. S. Carlson, M. A. Eberle, L. Kruglyak, and D. A. Nickerson, "Mapping complex disease loci in whole-genome association studies," *Nature*, vol. 429, no. 6990, pp. 446–452, 2004.

[47] W. S. Bush, S. M. Dudek, and M. D. Ritchie, "Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies," in *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, p. 368, NIH Public Access, 2009.

[48] . G. P. Consortium *et al.*, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, no. 7422, p. 56, 2012.

[49] J. P. Ioannidis, "Effect of formal statistical significance on the credibility of observational associations," *American journal of epidemiology*, vol. 168, no. 4, pp. 374–383, 2008.

[50] R. O'HARA, J. Cano, O. Ovaskainen, C. Teplitsky, and J. Alho, "Bayesian approaches in evolutionary quantitative genetics," *Journal of evolutionary biology*, vol. 21, no. 4, pp. 949–957, 2008.

[51] M. Stephens and D. J. Balding, "Bayesian statistical methods for genetic association studies," *Nature Reviews Genetics*, vol. 10, no. 10, p. 681, 2009.

[52] R. Fernando, A. Toosi, A. Wolc, D. Garrick, and J. Dekkers, "Application of whole-genome prediction methods for genome-wide association studies: a

bayesian approach," *Journal of Agricultural, Biological and Environmental Statistics*, vol. 22, no. 2, pp. 172–193, 2017.

[53] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly, "A new multipoint method for genome-wide association studies by imputation of genotypes," *Nature genetics*, vol. 39, no. 7, p. 906, 2007.

[54] M. Friedlander, A. Dobra, H. Massam, and L. Briollais, "Analyzing genome-wide association study data with the r package genmoss," *arXiv preprint arXiv:1611.07537*, 2016.

[55] Y. Guan and M. Stephens, "Practical issues in imputation-based association mapping," *PLoS genetics*, vol. 4, no. 12, p. e1000279, 2008.

[56] S. Winham, "Applications of multifactor dimensionality reduction to genome-wide data using the r package 'mdr'," in *Genome-Wide Association Studies and Genomic Prediction*, pp. 479–498, Springer, 2013.

[57] J. H. Moore, J. C. Gilbert, C.-T. Tsai, F.-T. Chiang, T. Holden, N. Barney, and B. C. White, "A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility," *Journal of theoretical biology*, vol. 241, no. 2, pp. 252–261, 2006.

[58] G. Abraham and M. Inouye, "Fast principal component analysis of large-scale genome-wide data," *PloS one*, vol. 9, no. 4, p. e93766, 2014.

[59] R. Bro and A. K. Smilde, "Principal component analysis," *Analytical Methods*, vol. 6, no. 9, pp. 2812–2831, 2014.

[60] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature genetics*, vol. 38, no. 8, p. 904, 2006.

[61] A. J. Izenman, "Linear discriminant analysis," in *Modern multivariate statistical techniques*, pp. 237–280, Springer, 2013.

[62] M. C. Stern, J. Lin, J. D. Figueroa, K. T. Kelsey, A. E. Kiltie, J.-M. Yuan, G. Matullo, T. Fletcher, S. Benhamou, J. A. Taylor, *et al.*, "Polymorphisms in dna repair genes, smoking, and bladder cancer risk: findings from the international consortium of bladder cancer," *Cancer research*, vol. 69, no. 17, pp. 6857–6864, 2009.

[63] S. Xu, "Theoretical basis of the beavis effect," *Genetics*, vol. 165, no. 4, pp. 2259–2268, 2003.

[64] B. Hayes, M. Goddard, *et al.*, "Prediction of total genetic value using genome-wide dense marker maps," *Genetics*, vol. 157, no. 4, pp. 1819–1829, 2001.

[65] L. Hosking, S. Lumsden, K. Lewis, A. Yeo, L. McCarthy, A. Bansal, J. Riley, I. Purvis, and C.-F. Xu, "Detection of genotyping errors by hardy–weinberg equilibrium testing," *European Journal of Human Genetics*, vol. 12, no. 5, p. 395, 2004.

[66] J. E. Wigginton, D. J. Cutler, and G. R. Abecasis, "A note on exact tests of hardy-weinberg equilibrium," *The American Journal of Human Genetics*, vol. 76, no. 5, pp. 887–893, 2005.

[67] P. Marjoram, A. Zubair, and S. Nuzhdin, "Post-gwas: where next? more samples, more snps or more biology?," *Heredity*, vol. 112, no. 1, p. 79, 2014.

[68] Z. Xu and J. A. Taylor, "Snpinfo: integrating gwas and candidate gene information into functional snp selection for genetic association studies," *Nucleic acids research*, vol. 37, no. suppl_2, pp. W600–W605, 2009.

[69] R. M. Nelson, M. Kierczak, and Ö. Carlborg, "Higher order interactions: detection of epistasis using machine learning and evolutionary computation," in *Genome-Wide Association Studies and Genomic Prediction*, pp. 499–518, Springer, 2013.

[70] S. Szymczak, J. M. Biernacka, H. J. Cordell, O. González-Recio, I. R. König, H. Zhang, and Y. V. Sun, "Machine learning in genome-wide association studies," *Genetic epidemiology*, vol. 33, no. S1, 2009.

[71] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," *Computational and structural biotechnology journal*, vol. 15, pp. 104–116, 2017.

[72] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[73] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[74] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.

[75] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[76] X. Chen and H. Ishwaran, "Random forests for genomic data analysis," *Genomics*, vol. 99, no. 6, pp. 323–329, 2012.

[77] Y. A. Meng, Y. Yu, L. A. Cupples, L. A. Farrer, and K. L. Lunetta, "Performance of random forest when snps are in linkage disequilibrium," *BMC bioinformatics*, vol. 10, no. 1, p. 78, 2009.

[78] W.-Y. Hwang, "Biological feature selection and disease gene identification using new stepwise random forests," *Industrial Engineering & Management Systems*, vol. 16, no. 1, p. 64, 2017.

[79] P. Mamoshina, A. Vieira, E. Putin, and A. Zhavoronkov, "Applications of deep learning in biomedicine," *Molecular pharmaceutics*, vol. 13, no. 5, pp. 1445–1454, 2016.

[80] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[81] S. Uppu, A. Krishna, and R. P. Gopalan, "Towards deep learning in genome-wide association interaction studies.," p. 20, 2016.

[82] S. Uppu, A. Krishna, and R. P. Gopalan, "A deep learning approach to detect snp interactions.," *JSW*, vol. 11, no. 10, pp. 965–975, 2016.

[83] P. Fergus, C. C. Montanez, B. Abdulaimma, P. Lisboa, and C. Chalmers, "Utilising deep learning and genome wide association studies for epistatic-driven preterm birth classification in african-american women," *arXiv preprint arXiv:1801.02977*, 2018.

[84] J. Kim, M. Kwak, and M. Bajaj, "Genetic prediction of type 2 diabetes using deep neural network," *Clinical genetics*, 2017.

[85] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," 2011.

[86] M. Szumilas, "Explaining odds ratios," *Journal of the Canadian academy of child and adolescent psychiatry*, vol. 19, no. 3, p. 227, 2010.

[87] F. Dorani and T. Hu, "Feature selection for detecting gene-gene interactions in genome-wide association studies," in *International Conference on the Applications of Evolutionary Computation*, pp. 33–46, Springer, 2018.