Ensemble Learning for Detecting Gene-Gene Interactions in Colorectal Cancer

by

© *Faramarz Dorani*

A thesis submitted to the

School of Graduate Studies

in partial fulfilment of the

requirements for the degree of

Master of *Science*

Department of *Computer Science*

Memorial University of Newfoundland

February 2018

St. John's Newfoundland

Abstract

The fundamental task of human genetics is to detect genetic variations that primarily contribute to a disease phenotype. The most popular method for understanding etiology of human inheritable diseases (e.g., cancer) is to utilize genome-wide association studies (GWAS). Colorectal cancer (CRC) is a common cause of deaths in developed countries; specifically, it has a high incidence rate in the province of Newfoundland and Labrador. Therefore, finding the affecting genetic factors associated with CRC can help better understand the disease in order to more effectively treat and prevent it. This study seeks to identify genetic variations associated with CRC using machine learning including feature selection and ensemble learning algorithms. In this study, we analyze a GWAS dataset on CRC collected from Newfoundland population. First, we perform quality control steps on the raw genetic data and prepare it for the machine learning methods. Second, we investigate six feature selection methods through a comparative study by applying them to a simulated dataset and CRC GWAS data. The best feature selection method, in terms of gene-gene interactions, is then used to choose a subset of more relevant features for the next step analysis. Subsequently, two ensemble algorithms, Random Forests and Gradient Boosting machine, are applied to the reduced data to identify significant interacting genetic markers associated with CRC. Last, the findings from machine learning methods are biologically validated using online databases and enrichment analysis tools. From the results of the ensemble algorithms, 44 significant genetic markers are detected in which 29 of them have corresponding genes in DNA. Among them, genes DCC, ALK and ITGA1 are previously found to be associated with CRC. In addition, there

are genes E2F3 and NID2, which have the potential of having association with CRC, because of their already known associations with other types of cancer. Moreover, the biological interpretations of these genes reveal biological pathways that may help predict the risk of the disease and better understand the etiology of the disease.

Keywords: machine learning, gene-gene interactions, colorectal cancer, feature selection, random forests, gradient boosting machine, GWAS, ensemble algorithms.

Acknowledgements

This work wouldn't been accomplished without constant help of my supervisor Dr. Ting Hu. During all of these two years, I bombarded her with my never-ending questions and she patiently guided me through them all. Therefore, my first thanks go to her. In addition, I appreciate Dr. Lourdes Penna-Castilo for helping me in answering different questions related to my thesis. Finally, I thank graduate students and my colleagues in the department of computer science for participating in informative talks that provided useful learning insights.

— Faramarz Dorani

Contents

List of Tables

List of Figures

List of Abbreviations

CART Classification and Regression Trees

CRC Colorectal Cancer

CV Cross-Validation

DAVID Database for Annotation, Visualization and Integrated Discovery

DNA Deoxyribonucleic acid

FS Feature Selection

GA Genetic Algorithms

GBM Gradient Boosting Machine

GI Gini Index

GO Gene Ontology

GWAS Genome-Wide Association Studies

GWA Genome-Wide Association

IG Information Gain

MAF Minor Allele Frequency

MDR Multifactor-Dimensionality Reduction

MI Mutual Information

ML Machine Learning

OOB Out-of-Bag

QC Quality Control

RF Random Forests

SNP Single Nucleotide Polymorphism

SURF Spatial Uniform ReliefF

SVM Support Vector Machine

TuRF Tuned ReliefF

VI Variable Importance

Chapter 1

Introduction

The fundamental task of human genetics is to detect genetic variations that primarily contribute to a disease phenotype. In these studies, the identification of genetic risk factors in inheritable and common diseases is the central goal [11]. There are two types of genetic inheritance: single gene inheritance also known as Mendelian inheritance, which is caused by mutations of DNA in a single gene, e.g., Cystic Fibrosis, and multi-factorial inheritance which is also called complex inheritance and caused by combination of environmental factors and multiple genes, e.g., heart disease and cancer.

In contrast to single-gene disorders, the approaches of study for complex or common diseases are not straightforward. Prior to the beginning of genomic studies, most of the experiments were performed based on familial linkage analysis on Mendelian diseases. However, this approach fails to reproduce for common diseases like hepatitis and cancer because of differences in the genetic architecture of common diseases and rare disorders [37, 44, 81]. To accomplish this purpose for common diseases,

researchers began to investigate a new research area: population-based genetic association studies, which deal with the investigation of the underlying genetic factors in a population to identify patterns of polymorphisms that vary systematically between individuals with different disease states [24]. In these studies, clinical genetic data of many individuals are collected and prepared for genotyping. Consecutively, a good deal of efforts are conducted to find the associations between genetic polymorphisms in the population and a measured trait or phenotype.

The naming convention for the population-based genetic association studies comes from the type of these studies. The term *population* refers to the individuals (or subjects) in the study who have no familial kinship. The term *association* refers to the mapping relationship between genetic variants and a trait (i.e., any effect or interaction between genetic variants an a trait.) The term *phenotype* is defined formally as a physical attribute or indicator of an individual's disease status (e.g., having or not having a disease). The terms trait, phenotype, and outcome are used broadly to refer to the same thing.

1.1 Genome-Wide Association Studies

The most promising type of population-based genetic association studies on common diseases is genome-wide association studies (GWAS or GWA studies) [24]. Based on the National Institutes of Health[1], a GWA study is defined as a study of common genetic variation across the entire human genome designed to identify genetic associations with observable traits [54]. The GWAS approach is an association study that

[1]https://www.nih.gov/

surveys most of the genome for identifying causal genetic variants in complex genetic diseases or traits [14, 36, 37].

In GWAS, the variations in DNA sequence from across the human genome are measured and analyzed using a sequencing technology such as next generation sequencing. The most common type of genetic variations in human genome are single nucleotide polymorphisms (SNPs, pronounced "snips") which are single variations in the DNA nucleotides among the population. GWAS are typically performed according to a case-control study design in which the cases are diseased and the controls are healthy individuals. In GWAS, the SNPs among a population of individuals (cases and controls) are genotyped and the corresponding genetic dataset is created [11]. GWAS data require large sample sizes and a large panel of genetic markers [11]. However, because of the costs associated with the data collection, GWAS data usually consists of hundreds of thousands of SNPs genotyped from hundreds to a few thousands of individuals. The dataset is then used by association analysis tools for investigating the relationship between genetic variants and disease trait.

1.2 Tools and Approaches in GWAS

Once a well-defined phenotype has been selected for a study population, and the genotypes are collected using an appropriate technique, the analysis of genetic data can begin [11]. Different approaches in GWAS are used to reveal the genetic risk factors in a disease. These approaches can be roughly divided into two categories: univariate analysis, and multivariate analysis.

1.2.1 Single-Locus Analysis

Most of the research in GWAS have been based on univariate techniques in which the relationship of one genetic factor to the disease phenotype is considered. These single-locus tests examine each SNP independently for association with the phenotype [11, 62]. The effect size (or penetrance) for any one variant is calculated and scored based on their significance of association with the disease phenotype. An example of a single-variable method is the *chi-square test of independence*, which measures the deviation from independence of genotypes and a phenotype under the null hypothesis.

The first successful single-variable study in GWAS was published in 2005 by Klein et. al. [44]. This case-control association study was designed for detecting genes involved in age-related macular degeneration (AMD). They performed single-maker associations testing and the results were two SNPs in gene CFH identified to be strongly associated with the disease.

Many single-SNP-based methods were used for some time, but had little success in detecting genetic risk factors [4, 81]. Single-variable methods produce some significant SNPs as primary contributors to disease state, but these SNPs only explain a small proportion of disease heritability and etiology [37, 81]. Moreover, single-variable methods and marginal testing analyses are less successful in finding associations because the causal SNPs are involved in an unknown genetic model (such as additive, dominant, or recessive), or may have epistatic interaction with other SNPs [4, 35]. The reason lies deep in the architecture of complex diseases, which are known to be caused by nonadditive interactions of multiple genetic variants or interaction of environmental factors and genetic variants that single-variable methods fail to detect.

1.2.2 Multi-Locus Analysis

Due to little success by utilizing the single-locus analysis methods [61], and because identified genetic variants from these methods explain only a small proportion of disease heritability, recent research have inclined toward using or developing multivariate approaches that examine interactions among genetic variants [4, 11, 25, 27]. It has been shown that it is not one genetic factor, rather interactions of multiple factors that contribute to susceptibility in complex diseases [62]. These interacting factors can be joint effect of multiple SNPs/genes, epistasis effect (e.g., SNP-SNP interactions and gene-gene interactions), and gene-environment interactions [41, 57].

Epistasis is an ubiquitous component of the genetic architecture of common human diseases [57]. It has been historically used to describe the phenomenon that the effect of a given gene on a phenotype can be dependent on one or more other genes. Indeed, it is an essential element for understanding the association between genetic and phenotypic variations [38, 62]. However, quantifying higher order epistasis is a challenging task due to both the computational complexity of enumerating all possible combinations in genome-wide data and the lack of efficient and effective methodologies. Epistasis, gene-gene interactions, and SNP-SNP interactions all convey the same concept in genome studies and they may be used interchangeably.

Multi-locus analysis methods are designed to find significant interactions among SNPs in GWAS data. However, the multi-locus analysis methods present numerous challenges regarding detecting interactions among SNPs. These challenges include: developing powerful statistical and computational methods to analyze genetic data, selecting appropriate genetic variables, and interpreting gene-gene interactions mod-

els [59].

1.2.3 Multi-Variate Analysis Approaches

There are three different approaches for doing a multi-locus analysis for variable interactions in GWAS data [94]. The first approach is exhaustive search methods e.g., multifactor-dimensionality reduction (MDR) which search through all combinations of underlying genetic factors [69]. Since most GWAS data have about one million genotyped SNPs, examining all pair-wise (or higher order) interactions between SNPs is a cost-prohibitive approach by most of the algorithms (even MDR). To resolve this issue, one approach is to use filtering methods to select only a subset of the most significant SNPs. MDR is preferable when the size of the feature set is relatively small (e.g., a few hundred).

A second approach is greedy search methods e.g., classification and regression trees (CART) [7]. These kinds of methods are able to detect interactions among variables and they are somehow preferable to exhaustive search methods, however, because of the greediness they may miss significant interactions among variables. An example of this approach is Random Forests (RF), which are composed of numerous CARTs built on the basis of random selection of variables [8]. The RF algorithm is capable of detecting interactions and has been used in many successful studies [81, 100].

A third approach is stochastic search methods e.g, evolutionary computing (EA) algorithms. These algorithms work based on the idea of natural selection and use a fitness (i.e., objective) function to find the optimum solution. Genetic algorithm (GA) is an example of EA algorithm which has been used in studies for dimensionality

reduction and multi-variate interaction detection [93].

1.2.4 Machine Learning

A machine learning (ML) algorithm learns through data to create a model that is used for future predictions [27]. Figure 1.1 shows diagram of a typical learning model in which the algorithm is trained based on training data and evaluated based on testing (or new) data. ML algorithms are capable of classification, regression, clustering, and feature ranking. ML methods have been the most commonly used approach in GWAS. Different ML approaches have been proposed and applied to GWAS data in order to model the relationship between SNPs and environmental factors to disease susceptibility for certain complex diseases [81]. Examples of ML methods which also been used in GWAS include: RF, Support Vector Machine (SVM), Naive Bayes classifier (NB), and Artificial Neural Networks (ANNs) [83].

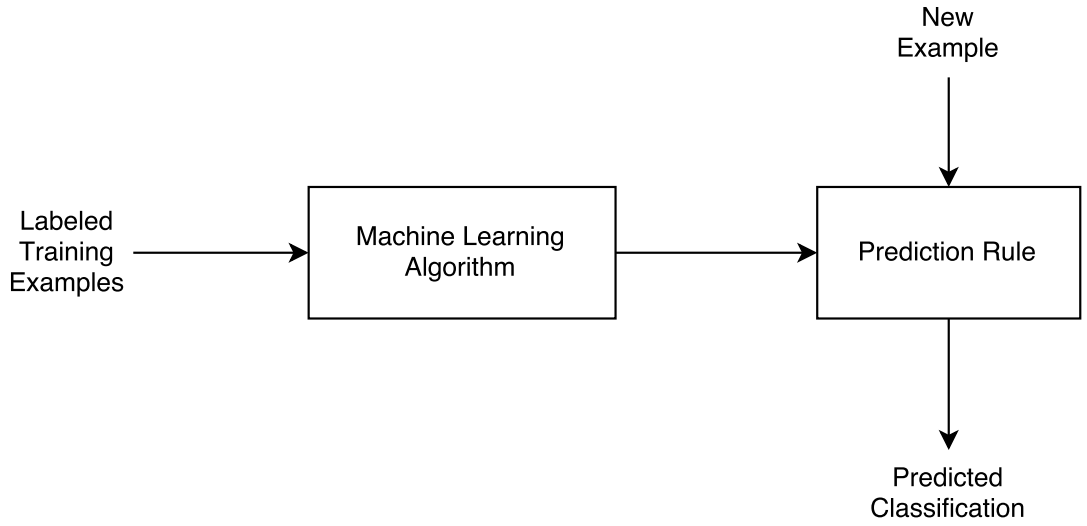

Figure 1.1: Diagram of a typical learning problem

1.3 Research Objectives

We report a whole-genome case-control association study for SNPs involved in colorectal cancer (CRC). What is unique about this study is the genetic data on which no gene-gene interaction analysis has been performed so far. Identifying gene-gene interactions in a genetic dataset is important because causality of common diseases, to a great extent, relies on the interactions among genetic variants. Thus, revealing interactions among genetic variants can help understand the etiology of the disease of interest.

We conduct a GWA study based on the workflow shown in Figure 1.2 [58]. In the beginning, we prepare the dataset by performing quality control on the genetic data to remove substandard samples, the ones with less genotypic information, and error-prone genetic variants, which contain erroneous genotypes. Subsequently, statistical analysis is conducted to evaluate the significance of the variants that might be used for dimensionality reduction. Simultaneously, we use dimensionality reduction methods such as filtering algorithms to reduce the size of the dataset to a moderate size which is applicable by computational methods. We then apply two ensemble algorithms, RF and Gradient Boosting Machine (GBM), to the reduced dataset to reveal interactions between SNPs. The ensemble methods have benefits over other methods because of their intrinsic multi-variable characteristics and their ability to detect interactions among variables. Subsequently, the results of computational/ML methods are interpreted to discover the most significant genetic factors in the disease. These significant genetic factors are biologically evaluated using genome knowledge databases in order to interpret as new discoveries.

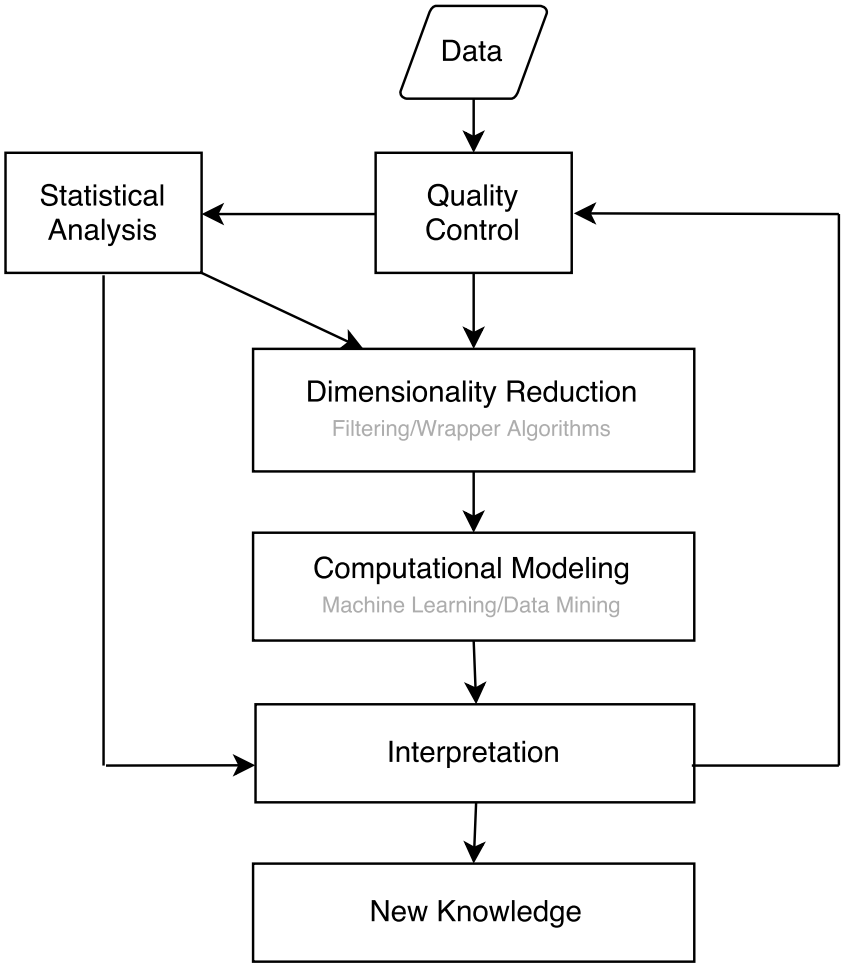

Figure 1.2: Workflow of this thesis research

Generally, conducting a GWA study involves four main steps as shown in the Figure 1.2. Data preparation and quality control, dimensionality reduction and feature selection, identifying significant genetic variants and/or detecting gene-gene interactions using computational methods, and conducting biological interpretations on the results of computational methods. All of the steps which have been conducted for the CRC dataset are explained in more detail in the below sections.

Chapter 2

Genome-Wide Association Studies for Colorectal Cancer

2.1 Background

Genomics is the study of genes and their functions. The objectives of genomic studies are to explore human genome to determine the functions of genes, find genetic variations and their significance in Deoxyribonucleic acid (DNA) sequence in development of diseases, and reveal the interactions of DNA and proteins. The first step in understanding human genome and how instructions are encoded in DNA (which lead to functions in humans) was the discovery of the sequence of the human genome accomplished at 2003 [87]. There are new technologies known as next-generation sequencing which have sped up the sequencing of all of a person's DNA. One method of next-generation sequencing is whole genome sequencing which determines the order of all the nucleotides (i.e., DNA building blocks) in an individual's DNA and can

determine variations in any part of the genome.

Human Genome Project (HGP) was an international research effort to sequence and map all of the genes - together known as the genome - of members of our species, Homo sapiens. Completed in April 2003, the HGP gave us the ability, for the first time, to read nature's complete genetic blueprint for building a human being.

2.1.1 International HapMap Project

Based on the *common variant/common disease* hypothesis, it is known that common diseases are caused by genetic variants that are common among people [11, 66]. That is, the heritability of common diseases can be explained to some extent by common genetic variations in the population. In this regard, the International HapMap Project was defined to describe and capture common genetic variations that are present in different human populations. In this project, the common genetic variations, or SNPs, across human genome are identified using different DNA sequencing techniques [30]. The SNP data are then used by researchers to reveal the relationship of the genetic variants and human diseases.

2.1.2 Genome-Wide Association Studies

In 2000, prior to the introduction of GWAS, the primary method of investigation for genotype-disease associations was through inheritance studies of genetic linkage in families. Linkage analysis is the attempt to statistically relate a transmission of an allele within families to the inheritance of a disease. This approach was proven to be highly useful towards single gene disorders or 'Mendelian' diseases. However, for

common and complex diseases, genetic linkage studies failed to produce good results [11, 37].

The completion of human genome sequence, which helps in SNP genotyping, and initiation of International HapMap Project have set stage for GWAS [37]. The current strategy for revealing the genetic basis of disease susceptibility is to carry out a GWA study [37, 58, 89]. GWAS are a new way of understanding human diseases. GWAS investigate genetic variation in human DNA to pinpoint genes that contribute to a particular disease risk. They are a promising approach to study complex and common diseases in which numerous genetic variations contribute to the disease status [14, 36, 37]. The goal of GWAS is to identify variants associated with the trait of interest using statistical and bioinformatic techniques [89]. This approach has been successful in identifying genetic variants that influence risks of complex diseases including cardiovascular [13, 56], autoimmune diseases [72], and cancer [22, 47, 55].

The chip-based microarray technology and, recently, next-generation sequencing have made it possible for GWAS to genotype $>$ one million SNPs [11]. Two primary platforms of SNP genotyping in GWAS are Illumina and Affymetrix, which utilize different underlying algorithms. After genotyping DNA sequences of individuals, the common genetic variants among them are determined. These common variations are known as SNPs, which are the most common type of genetic variation among people and occur more frequently in people with a particular disease than in people without a disease.

DNA is composed of building blocks called nucleotide (i.e., allele), of which there are four {A, T, C, G}. An SNP represents a difference in a single nucleotide and typically has two alleles demonstrating the possibilities of a base-pair at an SNP

locus [11]. There are three billion nucleotides within human DNA. SNPs occur once in every 300 nucleotides on average along the DNA sequence, meaning there are 10 million SNPs in the human genome. Most commonly, these variations are found in the DNA between genes and scarcely within genes. The small differences of SNPs may help predict a person's risk of particular diseases and response to certain medications.

Before SNP genotyping, the phenotype of interest (e.g., the type of disease) should be specified. Two groups of people, affected and unaffected, are then selected for genotyping. The affected are individuals having the disease (of interest) known as cases and the unaffected are healthy individuals known as controls. After preparing the SNP data of individuals, the genetic dataset of at least 1 million SNPs for about 1000 individuals is created. Subsequently, the dataset is analyzed by statistical or computational methods to unravel the most significant SNPs that contribute to the disease.

One main point to consider when doing a case-control association study is population stratification. The population stratification exists when the cases and controls are drawn from populations of different ancestry with different allele frequencies. It is important because it can result in false-positive results, because we might detect the population differences instead of loci associated with the disease [44]. Therefore, efforts need to be conducted to remove population stratification in the dataset.

2.1.3 Genome-Wide Association Studies Catalog

The National Human Genome Institute (NHGI) and European Bioinformatics Institute (EMBL-EBI) have been producing a catalog of all the eligible GWAS publications

since 2005 (`http://www.ebi.ac.uk/gwas`). For studies to be included in the GWAS catalog, certain criteria must be met. That is, the studies must include an analysis of $> 100,000$ SNPs, and SNP-trait associations must have a significance p-value $< 1.0 \times 10^{-5}$. As of November 13th 2017, the GWAS catalog contains 53,020 unique SNP-trait associations from 3,197 publications [53].

2.1.4 Colorectal Cancer

As stated in Chapter 1, we are conducting a GWA study on the CRC genetic data. CRC is the cancer of the large intestine (colon), which is 90% preventable if detected early. Factors that may increase the risk of colon cancer include: older age, family history of colon cancer, diabetes, obesity, smoking, and heavy use of alcohol. To prevent the risk of colorectal cancer there are recommendations such as: drinking milk, eating fruits, vegetables, and whole grains, exercising regularly, and stopping smoking[1].

CRC is the third most common type of cancer, which accounts for ∼10% of all cases of cancer [76]. CRC is a common cause of cancer deaths in developed countries with a high incidence rate in North America [98]. According to the Canadian Cancer Society (`http://www.cancer.ca/`), CRC is the second most commonly diagnosed cancer in Canada. It is the second leading cause of deaths from cancer in men and the third leading cause of deaths from cancer in women in Canada[2]. Estimations of Canadian colorectal cancer statistics for 2017 show that there would be approximately 14,900 male and 11,900 female cases, in which 5,100 of the male cases and 4,300 of

[1] https://www.mayoclinic.org/diseases-conditions/colon-cancer/symptoms-causes/syc-20353669

[2] http://www.cancer.ca/en/cancer-information/cancer-type/colorectal/statistics

the female cases result in deaths.

Based on Canadian Cancer Statistics, prostate cancer is the most frequently diagnosed type of cancer for men and breast cancer is the most frequent type of cancer in women in Newfoundland and Labrador. Newfoundland and Labrador also has the highest incidence rate of colorectal cancer for women as well. From 2017 statistics, 360 men and 270 women are estimated to be diagnosed with CRC. Moreover, 20.2% of the deaths from CRC will be individuals between the ages of 60-69 years and 27.7% of the deaths from CRC will be individuals between the ages of 70-79 years[3].

2.2 Methods

2.2.1 Dataset

For this study the CRC genetic data are collected from subjects within the province of Newfoundland and Labrador. The Colorectal Cancer Transdisciplinary (CORECT) consortium coordinated genotyping of data. Genotyping as part of the CORECT was conducted using a custom Affymetrix genome-wide platform (the Axiom CORECT Set) on two physical genotyping chips (pegs) for two datasets with ∼1.2 and ∼1.1 million SNPs [73]. The first dataset has 1,236,084 SNPs and 696 people with 200 cases and 496 controls in which the genotyping rate is 0.997134. The second dataset has 1,134,514 SNPs and 656 people with all cases and no controls with the genotyping rate of 0.888449. The genotyped SNPs in the datasets are not completely the same, rather, they have a small proportion of common SNPs and a high potential of overlap between subjects. Using PLINK[65], we merge these two datasets based on their common SNPs

[3]http://www.colorectal-cancer.ca/en/just-the-facts/colorectal/

resulting in the number of 265,195 variants and 1152 unique individuals. Among remaining phenotypes, 656 are cases and 496 are controls.

2.2.2 Quality Control

As for most population-based studies, the data need to be preprocessed before undergoing any further investigations. The preprocessing, which is performed as data quality assessments and control steps, is typically carried out during case-control association studies. Indeed, these steps are quite significant in the success of a case-control study and necessary before statistically testing for associations [3].

One important reason for the necessity of preprocessing is due to errors in genotype calling. During the genotyping of data (which is done by genotype calling algorithms) there is a possibility of occurring errors and missing values in the data. These errors could lead to an increase in false-positive and false-negative associations in case-control association studies [3]. In order to eliminate these issues and remove substandard samples and genetic markers, those assessments should be performed prior to any association analysis. Hence, quality control (QC) and missing value imputation (explained in next section) processes are conducted to prepare the data.

Using PLINK, a tool for handling genetic data [65], we perform quality assessments and control steps on the CRC data. PLINK provides commands for investigating the genetic data and performing the quality control steps. We undertake several quality control steps to remove individuals and markers with particularly high error rates. We take most of the steps from [3]. The detailed information on how to use PLINK can be accessed in [65]. There are usually two primary steps in QC: sample quality

and marker quality. Two main QC steps that were performed on the data are as follows:

1) Per-individual QC: per-individual QC or sample quality is a procedure to remove individuals with low-quality of genetic data. In this procedure, the individuals call rate, sample relatedness, and population stratification are investigated. Sample relatedness refers to the kinship of individuals in the population. It investigates if two individuals have a kinship relationship. Therefore, it computes the similarities between two individuals. Identical by descent (IBD) means that one individuals is identical to the other because they share the same DNA segment that they received through a parent. In other words, the proportion of loci where two individuals share zero, one or two alleles are referred as IBD.

The first step of QC is to remove samples with low-quality genotype information. The steps are consecutive such that each step uses the output data from the previous step. The steps are performed as follows:

a) First, we do a sex-check to identify individuals with problematic sex information.

```
plink --bfile raw-GWA-data --check-sex  --out raw-GWA-data
```

The `raw-GWA-data` is the PLINK's binary file of CRC data. This command produces sex information of individuals. We then find sex-discordant individuals and save them in the `fail-sexcheck-qc.txt` file. The command for removing those samples is:

```
plink --bfile raw-GWA-data --remove fail-sexcheck-qc.txt --make-bed
↪  --out clean-sexcheck-GWA-data
```

which removes individuals in the `fail-sexcheck-qc.txt` file and saves the result in the `clean-sexcheck-GWA-data` bed file.

b) The second step is removing sex chromosomes; that is, only chromosomes 1–22 are kept:

```
plink --bfile clean-sexcheck-GWA-data --chr 1-22 --make-bed --out
↪ clean-nosexchr-GWA-data
```

in which sex chromosomes are removed and the result is saved into the `clean-nosexchr-GWA-data` bed file.

c) The third step is removing samples with outlier missing genotypes. We use below command to produce missing genotype rate of samples:

```
plink --bfile clean-nosexchr-GWA-data --missing --out
↪ clean-nosexchr-GWA-data
```

then, the heterozygosity rate of samples are specified with the command below:

```
plink --bfile clean-nosexchr-GWA-data --het --out
↪ clean-nosexchr-GWA-data
```

Then, the samples with a missing genotype rate higher than 0.01 and a heterozygosity rate beyond $mean \pm 3sd$ (standard deviation) are identified and stored in `fail-imisshet-qc.txt` file. The following command removes these failing samples:

```
plink --bfile clean-nosexchr-GWA-data --remove fail-imisshet-qc.txt
↪ --make-bed --out clean-imisshet-GWA-data
```

Finally, those failing samples are removed and the result is saved into a bed file. So far, almost all of the error-prone samples are removed.

d) The fourth step is removing related individuals.

2) Per-marker QC: after removing low-quality samples, it is also important to remove sub-standard markers. The steps are performed in this regard are as follows:

a) The first step is removing low-quality markers using the following command:

```
plink --bfile clean-imisshet-GWA-data --geno 0.05 --maf 0.05 --hwe
↪ 0.0001 --make-bed --out clean-snp-GWA-data
```

in which SNPs with missing call rates exceeding 5%, a minor allele frequency (MAF) less than 5%, and with a Hardy-Weinberg equilibrium (HWE) greater than 0.0001 are removed.

b) The second step is removing markers with significant differences in the missing data rate between cases and controls. We identify missing data rates using the following command:

```
plink --bfile clean-snp-GWA-data --test-missing --out
↪ clean-snp-GWA-data
```

and then identify those SNPs with significant differences in the missing data and save them in the `fail-diffmiss-qc.txt` file. We then exclude those SNPs from the dataset using the following command:

```
plink --bfile clean-snp-GWA-data --exclude fail-diffmiss-qc.txt
↪ --make-bed --out clean-diffmiss-GWA-data
```

c) The third step is completed through linkage disequilibrium (LD) pruning, using the following command:

```
plink --bfile clean-diffmiss-GWA-data --indep-pairwise 2000 200 0.6
↪ --out ld-clean
```

which does a pairwise LD with window size of 2000 and r^2 of 0.6 and save those which pass the criteria in the `ld-clean.prune.in` file. We then only extract SNPs which are pruned in by LD using the following command:

```
plink --bfile clean-diffmiss-GWA-data --extract ld-clean.prune.in
↪ --make-bed --out clean-ld-GWA-data
```

d) The fourth step is creating a statistical recode dataset (0/1/2) using the following command:

```
plink --bfile clean-GWA-data --recode A --out clean-GWA-data
```

Allelic data are then recoded into genotype format, producing three categories for each SNP (0, 1 and 2 copies of the minor allele). Each SNP can be regarded as a bi-allelic variable, i.e., it has two different variations, with the common allele among a population called the *reference* and the other called *variant*. Given the fact that human chromosomes are paired, three categorical values are usually used to code for each SNP, i.e., 0 for homozygous reference, 1 for heterozygous variant, and 2 for homozygous variant. The controls and cases are assigned to class labels to 1 and 2 respectively. The final dataset is created and stored in the `clean-GWA-data` file which would be used hereafter for subsequent analysis.

2.2.3 Imputation

One important note we notice about the cleaned CRC dataset is the imbalanced class labels. When building ML models, imbalanced class labels in the dataset usually inject a bias into the model. That is, machine learning models tend to make predictions toward the class with higher frequency. In the cleaned CRC dataset, the number of 626 cases is high in comparison to the 472 controls. These imbalanced labels cause the prediction models to predict all of the labels of test data as one class –which is *case* in this situation. Therefore, we make the dataset balanced by removing cases

with numerous missing values rather than removing (more) low-quality SNPs. We count the number of missing values for each case and remove the ones which have 10% missing values. In this way, we would have a clean dataset with the balanced class labels.

After balancing the dataset, we impute the missing values that are not too many with an algorithm. The imputation algorithm replaces missing values with the most appropriate values from the dataset. That is, the most frequent value in each SNP is found and put into the missing places. The reason for choosing this imputation method is that the CRC dataset after QC and balancing do not have enormous missing values. In contrast, the frequency of missing values in SNPs are less than 10% which is not significant.

2.3 Results

The consecutive runs of the PLINK commands remove low-quality SNPs and samples from the CRC dataset. From the per-individual QC steps, in the execution of step (a) 11 samples with inconsistent of sex information, in step (b) no samples, and in step (c) 26 samples with outlier missing genotypes are removed from the dataset. From the per-marker QC steps, in the execution of step (a) 20,693 low-quality SNPs with a genotype rate less than 5%, a minor allele frequency less than 5%, and a Hardy-Weinberg equilibrium greater than 0.0001, in step (b) 1,257 SNPs with significant differences in the missing data rate between cases and controls, and in step (c) 47,046 SNPs which are in linkage disequilibrium with each other are removed. The quality control steps resulted in a dataset with 190,142 SNPs for 1,098 individuals.

As stated in the imputation section, the cases with more than 10% missing values are removed, that result in the number of 944 samples with 472 cases and 472 controls. Subsequently, we again remove low-quality SNPs based on a threshold by counting the number of missing values for each SNP. We remove SNPs which have more than about 1% (10 out of 944 samples) missing values in the samples. Table 2.1 shows summary information of original and clean CRC dataset after all quality control steps and imputation. At the last stage, we replace the missing values with the most frequent value for each SNP. The final dataset would have 186,251 SNPs for each 944 samples.

Table 2.1: CRC dataset information

Stage	SNPs	Samples	Cases	Controls
Before QC	265,195	1,152	656	496
After QC	190,142	1,098	626	472
After Imputation	186,251	944	472	472

Chapter 3

Feature Selection

3.1 Background

It is a challenging task to analyze high dimensional SNP data for GWAS. The number of variables, i.e., SNPs, brings an extensive computational burden for informatics methods [58]. Moreover, in the studies of common human diseases, it has been accepted that the non-additive effects of multiple interacting genetic variables play an important role explaining the risk of a disease [16]. The traditional one-gene-at-a time strategies likely overlook important interacting genes that have moderate individual effects. Therefore, powerful data mining and machine learning methods are needed in order to examine multiple variables at a time and to search for gene-gene interactions that contribute to a disease. A GWAS dataset with a million variables can be prohibitive for the application of any machine learning algorithms for detecting gene-gene interactions, since enumerating all possible combinations of variables is impossible. In addition, many of those variables can be redundant or irrelevant for the

disease under consideration. Thus the selection of a subset of relevant and potential variables to be included in the subsequent analysis, i.e., *feature selection*, is usually needed [58].

Feature selection (FS) is frequently used as a pre-processing step in machine learning when the original data contain noisy or irrelevant features that could compromise the prediction power of learning algorithms [97]. FS methods choose only a subset of the most important features, and thus reduce the dimensionality of the data, speed up the learning process, simplify the learned model, and improve the prediction performance [19, 33].

3.1.1 What is Feature Selection?

Feature selection is referred to as the process of selecting a subset of features from a feature set in a dataset [71]. This process is usually performed before classification modeling. Actually, it is an important scientific requirement for a classifier when the number of variables is large compared with the number of subjects [27]. Dimensionality reduction or FS is worthy in the sense that they reduce the computational complexity of future classification models by providing a fewer number of features. Sometimes, it provides more reliable data by removing noise variables [71, 33]. Feature selection involves two main objectives, i.e., to maximize the prediction accuracy and to minimize the number of features.

There are four basic criteria that should be considered when designing an FS method [5]. *Direction of search* is the determination of starting point from which to start searching. The examples are forward selection or backward elimination.

Search space specifies the organization of the search. One way is exhaustive search of all 2^a possible subsets of a attributes. Another wise approach is greedy search to traverse the space. *Evaluation of subset of features* is also important in assessing the significance of selected features. This can be done by measuring accuracy on the training or test sets. *Halting of search* is the determination of a termination criterion to decided when to stop the search. This criterion could be a specified number of attributes, the accuracy of classifier or combination of both.

3.1.2 Types of Feature Selection

Generally, FS methods can be divided into three categories. *Filter approach* separates feature selection from classifier such that at first a subset of features are filtered in and then fed into the classifier. In other words, the learning algorithm plays no role in selecting attributes. This approach could also rank attributes based on their significance. Moreover, in comparison to its companions this method is relatively fast. However, the downside of this approach is the accuracy because it is not being evaluated by classifier [26]. An example of this approach is Kira and Rendell's Relief algorithm which uses a complex feature evaluation function [43].

Wrapper approach iteratively evaluates performance of selected features until certain accuracy is reached. In other words, a wrapper algorithm searches through the feature space using a filter method, but feature evaluation is performed via a classifier. The accuracy of classifier on some training data is used as the metric of evaluation. This approach is more useful when the size of dataset is small and classification is of great importance. The major disadvantage of wrapper algorithm is computational

cost because it has to call the classifier whenever evaluating the feature set. Therefore, improvements on speeding evaluation is needed in this method such as using greedy or stochastic search rather than deterministic search [5, 26, 33, 71].

Embedded approach embeds feature selection within classifier. These methods could be helpful in detecting correlation among variables. Examples of embedded algorithms are Decision trees, Naive Bayes, and penalized methods having penalty functions such as Lasso and Elastic Net [52, 71].

Of those three, filter approaches are often used in bioinformatics studies given the fact that they can easily scale to very high-dimensional data, that they are computationally simple and fast, and that they are independent of the classification algorithm [71]. In addition, most of the used FS algorithms in GWAS are filter-based methods, since filtering methods outperform other methods in large-scale problems [39].

There are other dimensionality reduction techniques such as principal components analysis and partial least squares in which the original input variables are transformed into a new input variables. This could be helpful for classification problems but not useful in biomedical implications, since the original input variables are deformed and not easily accessible [52]. Therefore, we will not include these methods in our study. Furthermore, there is another type of feature selection or dimensionality reduction methods called hybrid methods such as sparse principal component analysis which is combination of feature selection and dimensionality reduction. However, we will not be using these methods as well. Feature selection methods are preferable to dimensionality reduction methods in bioinformatics because FS methods do not change the behavior of genetic variants.

3.1.3 Feature Selection in Bioinformatics

There have been studies investigating the performance of FS methods on high dimensional datasets in bioinformatics. Hua et al. [39] evaluated the performance of several filter and wrapper feature selection methods on both synthetic and real gene-expression microarrays data. The size of datasets was 20,000 features (i.e. genes) and the largest sample size was 180. They used a two-stage feature selection strategy where filter methods were applied before training the classifier to remove non-informative attributes and then wrapper methods were used to refine feature set.

Shah and Kusiak [75] used genetic algorithms (GA) to search for the best subset of SNPs in a dataset with 172 SNPs. After selecting the best subset of SNPs by GA, the subset is evaluated by a baseline classifier to compare the performance when whole feature set is used.

Wu et al. proposed an SNP selection and classification approach on GWAS data based on RF. In the proposed stratified random forest (SRF) method, SNPs are spread into groups based on the significance of their informativeness computed based on information gain. Then, using resampled data, the CART trees are grown by selection of a number of SNPs from each group. The method has been tested on Parkinson and Alzheimer case-control data and compared to other methods such as original RF (with different parameter values), and SVM.

Bermingham et al. investigated performance of five feature selection methods (4 supervised and 1 unsupervised) on two classification methods: a mixed model (G-BLUP) and a Bayesian (Bayes C). Three of the supervised feature selection methods were based on p-value rankings of SNPs associations in dataset, and the fourth one

was based on partitioning the SNPs into haplotype blocks and the p-value of intra-block SNPs covariates. The unsupervised feature selection method was based on random selection of different number of SNPs that are evenly spaced from each other. The methods were tested on two genome-wide SNP datasets, Croatian and UK, with 2,186 and 810 individuals respectively.

Numerous mutual information (MI) feature selection methods have been proposed in the last 20 years. MI is a measure of statistical independence between two random variables. It is a measure of the amount of information that one random variable has about another variable. Brown et al. proposed a framework for information-theoretic feature selection methods in which to select the smallest feature subset having the highest MI [9].

3.2 Methods

Most existing studies used the classification accuracy as the indicator for feature selection performance. The contribution of a feature to a phenotypic outcome could be its individual main effect or its interacting effect combined with other features. Using the overall classification accuracy was not able to distinguish the interaction effects of multiple variables and the individual main effects.

In our study, we focus on searching for features (SNPs) that have strong associations with the disease outcome in terms of gene-gene interactions. This differentiates our work from many existing studies that mostly focus on SNPs with high main-effects. We apply information gain to quantify the pair-wise synergy of SNPs and use that to evaluate various feature selection methods in order to identify the ones

that can find subsets of SNPs with high synergistic effects on the disease status. We investigate six most popular filter algorithms, and test them on both simulated and real GWAS datasets. Our findings can be helpful for the recommendation of feature selection methods for detecting gene-gene interactions in GWAS.

In this section, we first discuss the data that will be used in this study, which include a simulated and a real population-based GWAS datasets. Then we introduce the information gain measure that will be employed as the quantification of the synergistic interaction effect of pairs of SNPs. Last, we present the six feature selection algorithms that will be investigated and compared.

3.2.1 Simulated Data

Having an understanding of performance of FS and statistical methods on real genetic data is not straightforward. In some studies, the performance of FS methods is evaluated on a simulated data to have a grasp on capability of these methods in identifying important SNPs. For this purpose, we created a simulated data of SNPs with the genetic characteristics similar to the CRC data.

For this study, we use a simulated genetic association dataset generated by the genetic architecture model emulator for testing and evaluating software (GAMETES) [85, 84]. GAMETES is a fast algorithm for generating simulation data of complex genetic models. Particularly, in addition to additive models, GAMETES is specialized for generating pure interaction models, i.e., interacting features without the existence of any main effects. Each n-locus model is generated deterministically, based on a set of random parameters and specified values of heritability, minor allele frequencies, and

population disease prevalence. Since we focus on pairwise SNP interactions, we use GAMETES to generate a population of 500 samples with half being cases and half being controls. The dataset has 1000 SNPs, where 15 pairs are two-locus interacting models with a minor allele frequency of 0.2 and another 970 are random SNPs. We set the heritability to 0.2 and population prevalence to 0.5.

3.2.2 Quantification of Interactions Using Information Gain

Information theoretic measures such as entropy and mutual information [17] quantify the uncertainty of single random variables and the dependence of two variables, and have seen increasing applications in genetic association studies [23, 48, 38]. In such a context, the *entropy* $H(C)$ of the disease class C measures the unpredictability of the disease, and the conditional entropy $H(C|A)$ measures the uncertainty of C given the knowledge of SNP A. Subtracting $H(C|A)$ from $H(C)$ gives the *mutual information* of A and C, and is the reduction in the uncertainty of the class C due to the knowledge about SNP A's genotype, defined as

$$I(A;C) = H(C) - H(C|A). \tag{3.1}$$

Mutual information $I(A;C)$ essentially captures the main effect of SNP A on the disease status C.

When two SNPs, A and B, are considered, mutual information $I(A,B;C)$ measures how much the disease status C can be explained by combining both A and B. The *information gain* $IG(A;B;C)$, calculated as

$$IG(A;B;C) = I(A,B;C) - I(A;C) - I(B;C), \tag{3.2}$$

is the information gained about the class C from the genotypes of SNPs A and B considered together minus that from each of these SNPs considered separately. In brief, $IG(A;B;C)$ measures the amount of synergetic influence SNPs A and B have on class C. Thus, information gain IG can be used to evaluate the pairwise interaction effect between two SNPs in association with the disease.

3.2.3 Feature Selection Methods

We choose six most widely used feature selection algorithms in our comparative study, and investigate their performance on searching variables that contribute to the disease in terms of gene-gene interactions. These six feature selection algorithms include three uni-variate approaches, chi-square, logistic regression, and odds ratio, and three Relief-based algorithms, ReliefF, TuRF, and SURF. They will be applied to both simulated and real GWAS datasets and provide rankings of all the SNPs in the data.

Chi-square: The chi-square (χ^2) test of independence [96] is commonly used in human genetics and genetic epidemiology [58] for categorical data. A χ^2 test estimates how likely different alleles of a SNP can differentiate the disease status. It is a very efficient filtering method for assessing the independent effect of individual SNPs on disease susceptibility.

Logistic regression: Logistic regression measures the relationship between the categorical outcome and multiple independent variables by estimating probabilities using a logistic function. A linear relationship between variables and the categorical outcome is usually assumed, and a coefficient is estimated for each variable when such a linear relationship is trained to best predict the outcome. The variable coefficient

can then be used as a quantification of the importance of each variable.

Odds-ratio: Odds ratio (OR) is the most commonly used statistic in case-control studies. It measures the association between an exposure (e.g., health characteristic) and an outcome (e.g., disease status). The OR represents the odds that a disease status will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure [80].

ReliefF: Relief is able to detect complex attribute dependencies even in the absence of main effects [43]. It estimates the quality of attributes using a nearest-neighbor algorithm. While Relief uses, for each individual, a single nearest neighbor in each class, ReliefF, a variant of Relief, uses multiple, usually 10, nearest neighbors, and thus is more robust when a dataset contains noise [45, 70]. The basic idea of Relief-based algorithms is to draw instances at random, compute their nearest neighbors, and adjust a feature weighting vector to give more weights to features that discriminate the instance from its neighbors of different classes. Comparing to uni-variate feature selection algorithms, ReliefF is able to capture attribute interactions because it selects nearest neighbors using the entire vector of values across all attributes [58, 70].

Tuned ReliefF (TuRF): It is an extension of ReliefF specifically for large-scale genetics data [60]. This method systematically and iteratively removes attributes that have low-quality estimates so that the remaining attributes can be re-estimated more accurately. It improves the estimation of weights in noisy data but does not fundamentally change the underlying ReliefF algorithm. It is useful when data contain a large number of non-relevant SNPs. It is also more computationally intense because of the iterative process of removing attributes.

Spatially Uniform ReliefF (SURF): SURF is also an extension of the ReliefF algorithm [32]. It incorporates the spatial information when assesses neighbors. Instead of using a fixed number of neighbors as the threshold in ReliefF, SURF uses a fixed distance threshold for choosing neighbors. It is reported to be able to improve the sensitivity detecting small interaction effects.

3.3 Results

3.3.1 Feature Selection Algorithms on the Simulated Data

First, we apply all six feature selection algorithms to the simulated dataset that contains 30 known SNPs with pairwise interactions and 970 random SNPs. The chi-square, odd-ratio, ReliefF, TuRF, and SURF algorithms are implemented using the multifactor dimensionality reduction (MDR) software with default parameter settings [69]. Logistic regression is implemented using the Python *scikit-learn* package [64].

Each algorithm yields a ranking of all 1000 SNPs. Table 3.1 shows the statistics of the ranking scores of those 30 known SNPs by each feature selection algorithm. We see that TuRF has both the best mean and median rankings among all the methods, and the differences are significant. ReliefF performs the second best, followed by SURF.

Table 3.1: Ranking of the 30 known interacting SNPs by feature selection algorithms.

	Logit	χ^2	OR	ReliefF	TuRF	SURF
Mean	549.16	548.30	444.10	202.63	**166.96**	233.16
SD	277.99	267.18	287.04	201.74	259.74	212.13
Median	617.50	536.50	346.50	130.00	**21.50**	183.50

Figure 3.1 shows the recall-at-k for all six feature selection algorithms. The y-axis shows the fraction of those 30 known SNPs detected by the top k SNPs ranked by each feature selection algorithm. We can see that for all values of k, TuRF has the highest recalls. In addition, all three Relief-based algorithms outperform the other methods.

Figure 3.2 shows the distributions of the ranking of those 30 known interacting SNPs using different feature selection algorithms. The x-axis is the rank of SNPs and the y-axis is the density. Again, TuRF has the highest density around high ranks, meaning that it produces the highest ranks for those 30 known SNPs. SURF and ReliefF also have better ranking performance comparing to the other three methods. Odds-ratio, logistic regression, and chi-square have flat distributions across the entire rank range, which indicates their inability to identify those 30 interacting SNPs.

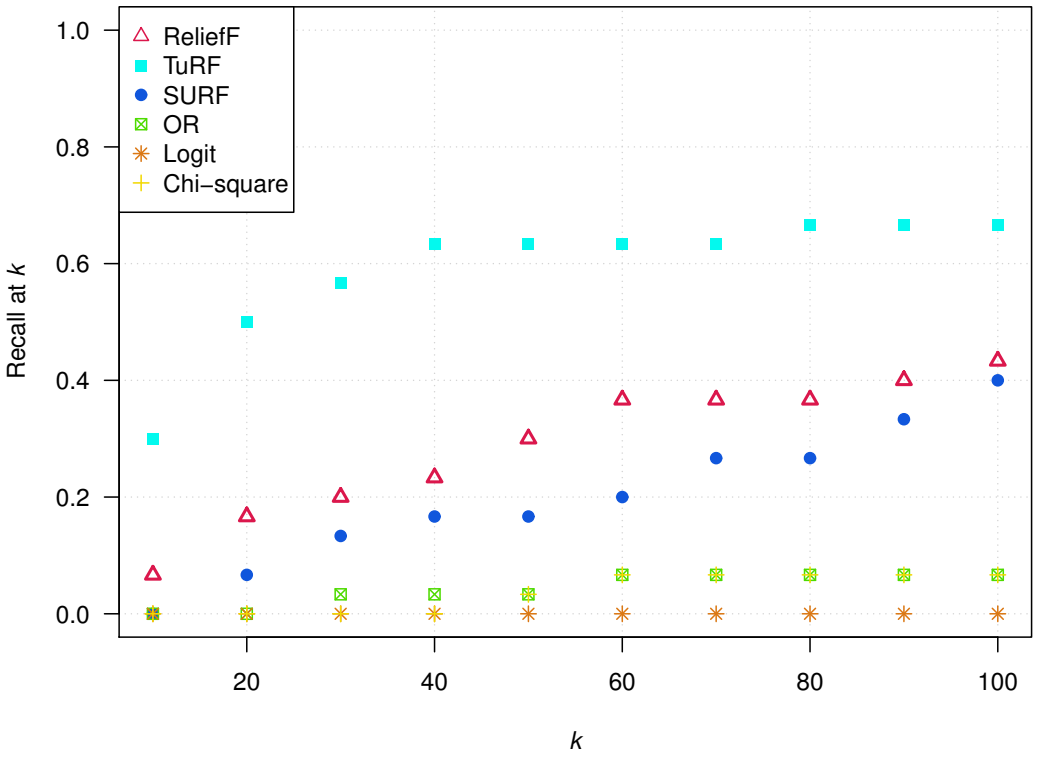

Figure 3.1: Diagram of recall-at-k for six feature selection algorithms applied to the simulated dataset. Recall-at-k is the fraction of the 30 known interacting SNPs detected by the top k ranked SNPs using each feature selection algorithm.

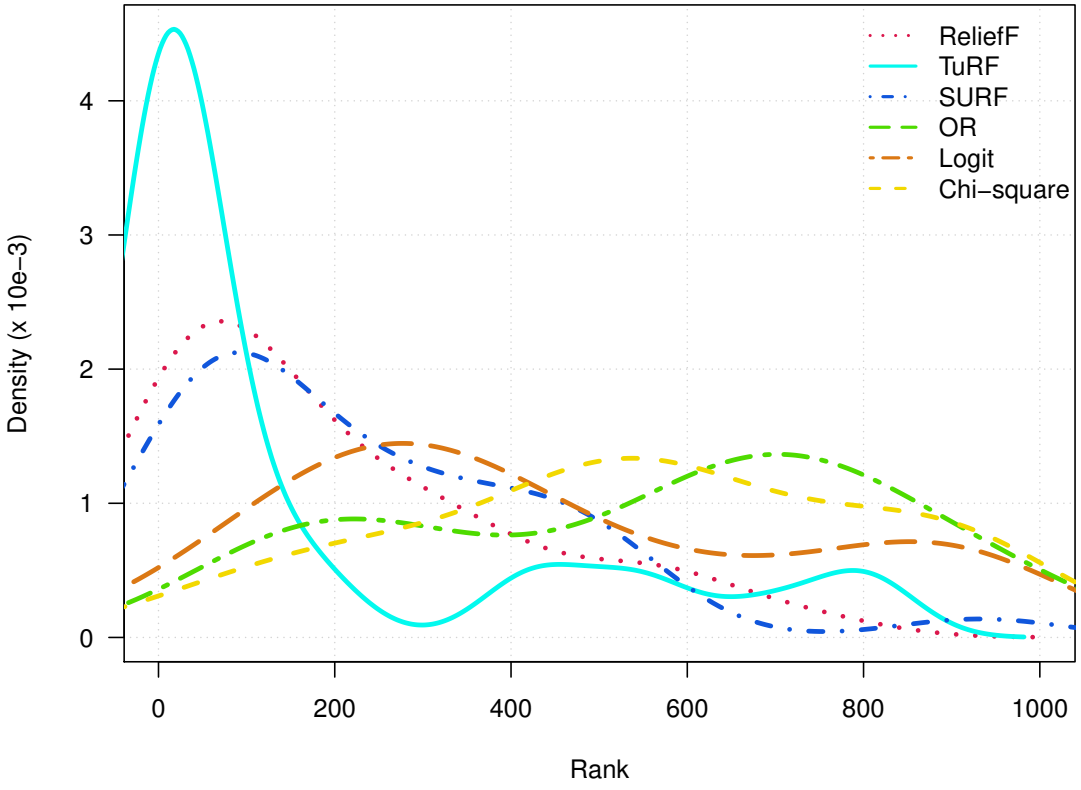

Figure 3.2: Density of the rankings of the 30 known interaction SNPs using different feature selection algorithms on the simulated dataset.

3.3.2 Feature Selection Algorithms on the CRC Data

We then compare the performance of those six feature selection algorithms using the CRC GWAS dataset. The CRC GWAS dataset is processed using PLINK software [65]. PLINK can conduct some fundamental association tests by comparing allele frequencies of SNPs between cases and controls. We use the command `--assoc` to compute chi-square and odds-ratio scores for each SNP, and the command

`--logistic` for logistic regression analysis. Again, we used the MDR software [69] to implement ReliefF, TuRF, and SURF algorithms.

Each feature selection algorithm generates a ranking of all the 186,251 SNPs in the dataset. For detecting gene-gene interactions, exhaustive enumeration of all possible combinations of SNPs is usually considered. Even for pairwise interactions, the total number of possible pairs $\binom{n}{2}$ grows fast with the number of SNPs n. Therefore, we can only consider a moderate subset of SNPs for interaction analysis, and we use the rankings estimated using feature selection algorithms to filter those potentially more important SNPs. We choose the subset of the top 10,000 SNPs by each feature selection algorithm. Then, for the six subsets of filtered 10,000 SNPs, we evaluate their pairwise interactions separately using the information gain (IG) measure.

Table 3.2 shows the maximum, minimum, mean, standard deviation, and median values of the information gain calculated using all $\binom{10{,}000}{2}$ pairs of the 10,000 SNPs filtered by the six feature selection algorithms. As we can see, ReliefF finds the SNP pair with the highest interaction strength, and TuRF has the best overall distribution.

Table 3.2: Statistics of the information gain values of all $\binom{10,000}{2}$ SNP pairs filtered by each feature selection algorithm ($\times 10^{-3}$).

	Logit	χ^2	OR	ReliefF	TuRF	SURF
Max	27.4	27.6	27.4	**30.2**	28.9	28.2
Min	-4	-5.1	-4	-3.2	-2.9	-5.7
Mean	2.760	3.047	2.776	3.190	**3.191**	3.056
SD	2.117	2.221	2.120	2.243	2.251	2.224
Median	2.3	2.6	2.3	2.7	2.7	2.6

Fig. 3.3 shows the distribution of the interaction strength IG of all $\binom{10,000}{2}$ pairs of SNPs selected by each feature selection algorithm. We see that the distributions of ReliefF and TuRF have overall more SNP pairs with higher IG values. The distributions of SURF and chi-square are comparable, and logistic regression and odds ratio have the lowest overall IG values.

The significance of the IG value of each pair of SNPs can be assessed using permutation testing. For each permutation, we randomly shuffle the case/control labels of all the samples in the data in order to remove the association between the genotypes of SNPs and the disease status. Repeating such a permutation multiple times generates a null distribution of what can be observed by chance. For each permuted dataset, we compute the IG value of each pair of SNPs. In this study, we perform a 100-fold permutation test rather than 1000 permutations because of the computational (space and time) complexity imposed by calculating higher order permutations. The significance level (p-value) of the IG of each SNP pair can be assessed by comparing the IG

value of the pair calculated using the real dataset to the IG values calculated using the 100 permuted datasets (see Algorithm 1).

Algorithm 1 Permutation testing algorithm

1: **procedure** COMPUTEPVALUE
2: $D \leftarrow$ *original dataset*
3: $n \leftarrow$ *number of permutations*
4: $m \leftarrow$ *number of SNP pairs*
5: $C \leftarrow$ *counter for each SNP pair*
6: $i \leftarrow 1$
7: $j \leftarrow 1$
8: **while** $i < n$ **do**
9: *permute the dataset* D
10: **while** $j < m$ **do**
11: *calculate* $IG_i^{permute}$ *for the* j*-th SNP pair*
12: *increase* C_j *by 1 if* $IG_i^{permute}$ *is greater than the observed* IG_j
13: *calculate the significance level* p_j *for each SNP pair* j *as* $\frac{C_j}{n}$

We apply permutation testing to all six subsets of $\binom{10,000}{2}$ pairs of SNPs selected by each feature selection algorithm, such that their significance level p-values can be assessed. Fig. 3.4 shows the number of SNP pairs that pass two different p-value thresholds, 0.01 and 0.05. TuRF has more SNP pairs with significant interaction strength using both thresholds. All three Relief-based algorithms have higher numbers of significant SNP pairs than the other three methods. Logistic regression and odds

ratio find the least numbers of significant interacting SNP pairs.

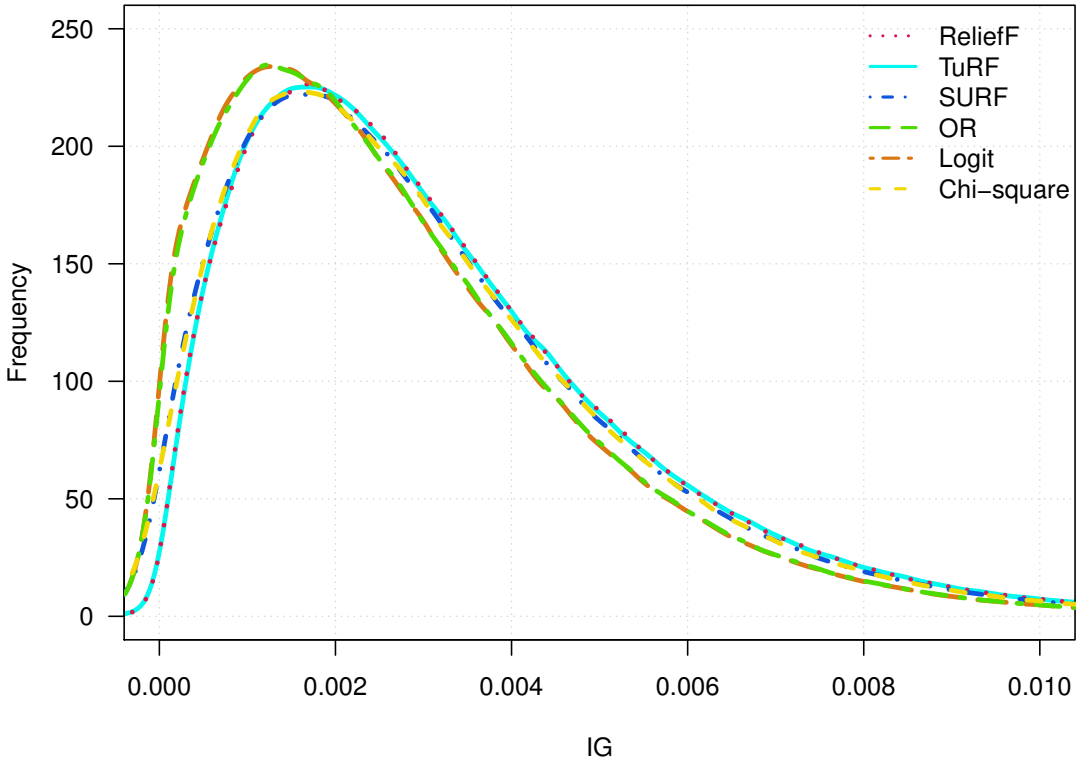

Figure 3.3: Distribution of the information gain (IG) values of all pairs of filtered 10,000 SNPs by each feature selection algorithm.

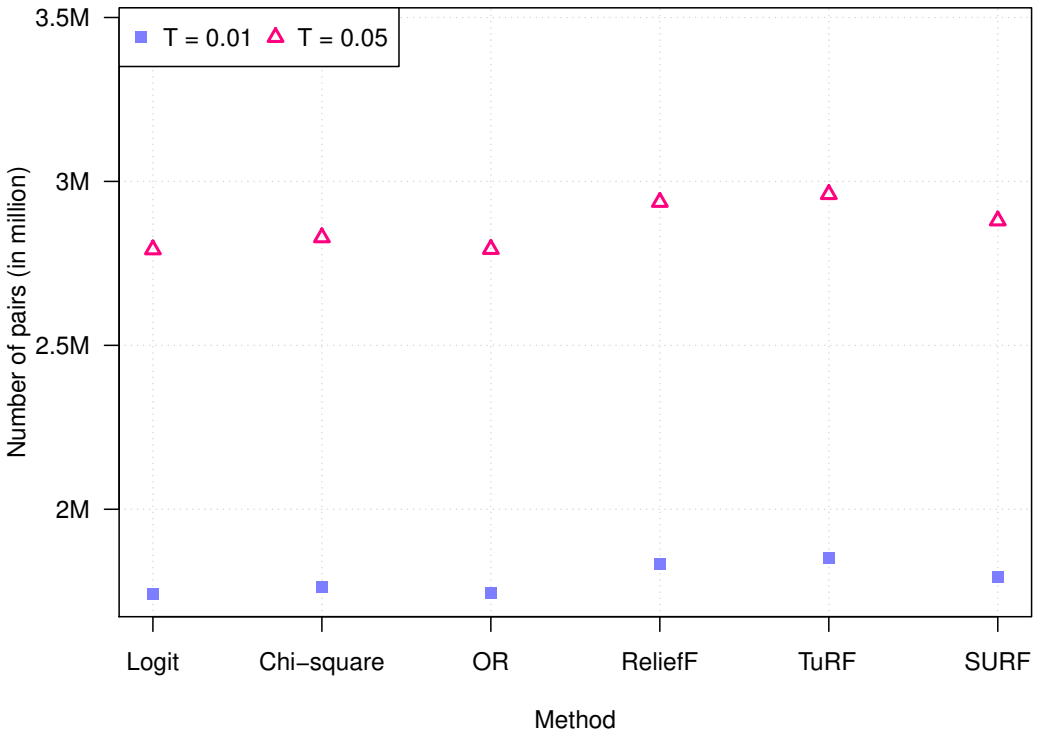

Figure 3.4: The number of SNP pairs with significant interaction strengths using p-value cutoff T. Red triangles show the results using cutoff $p \leq 0.01$, and blue squares show the results with cutoff $p \leq 0.05$.

3.3.3 Applying TuRF Feature Selection Method

TuRF feature selection outperformed other FS methods in terms of detecting significant interactions among SNPs. Therefore, we use the TuRF feature selection method to give scores to SNPs based on their significance in associating with the disease status. The next step is then to choose a reasonable threshold to filter in the most significant SNPs produced by the TuRF method. We specify a threshold, which

gives an appropriate number of the most significant SNPs, as long as the dataset created from those SNPs would not produce computational burden for the ML methods. Based on the histogram of SNPs' score shown in Figure 3.5, the threshold $mean+3sd$ is chosen, in which $mean$ is the average and sd is the standard deviation of all of the SNP scores, to obtain the most significant SNPs. A total number of 2,798 SNPs are selected for the subsequent machine learning analysis.

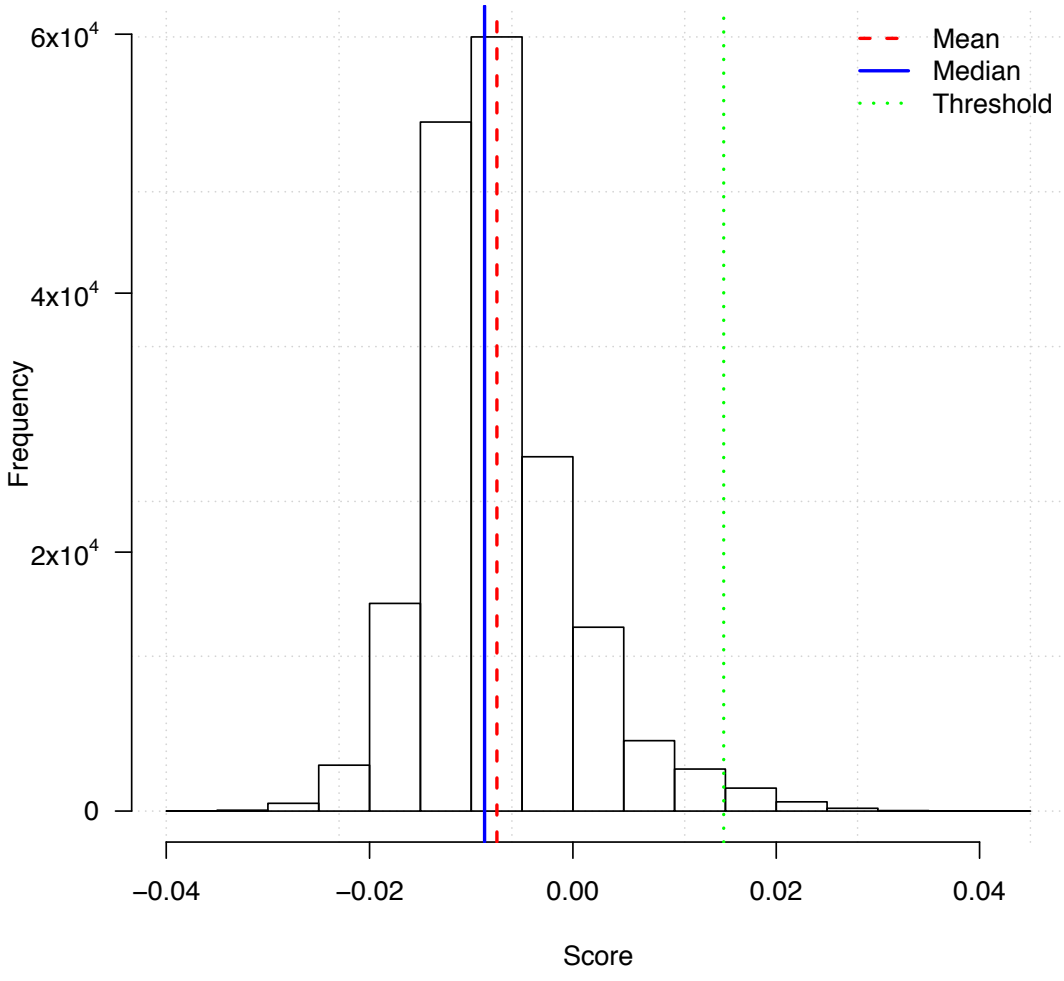

Figure 3.5: Histogram of SNP scores by the TuRF method

3.3.4 Discussion

In this chapter, we investigated the performance of six widely used feature selection algorithms for detecting potentially interacting single nucleotide polymorphisms (SNPs) for GWAS. We used both a simulated and real genetic datasets. We adopted information gain as a measure for quantifying pairwise interaction strength of SNPs in order to evaluate the filtering performance of those six feature selection algorithms. Among the investigated feature selection methods, three are single variable feature scoring methods. That is, they only consider individual main effects of SNP on the disease status. Three other methods are extensions of the Relief algorithm which is a multivariate feature selection algorithm.

For the simulated dataset, we generated a population-based dataset with 1000 SNPs including 15 pairs of interacting SNPs and 970 random ones. We applied all six feature selection algorithms to rank those 1000 SNPs and look into the recall-at-k of detecting those 30 known interacting SNPs. The TuRF algorithm has the highest recall-at-k for all k values, followed by ReliefF and SURF. All three Relief-based algorithms perform better than odds ratio, logistic regression, and chi-square.

We also tested the feature selection algorithms using a real GWAS dataset on colorectal cancer (CRC). We used information gain to quantify pairwise interaction strength of SNPs in order to evaluate the filtering performance of the feature selection algorithms. We chose 10,000 top-ranked SNPs by each feature selection algorithm and applied information gain measure and permutation testing to compute the interaction strengths and their significance levels of all pairs of SNPs. We found that TuRF again was able to filter more significant interacting SNPs than the rest of the feature

selection algorithms. All three Relief-based algorithms outperformed the other three methods.

TuRF and ReliefF had comparable performance on the application to the real CRC dataset. By looking at their top 10,000 SNPs, we saw that only 1474 were overlapped. That is, only 14% of their top 10K SNPs are the same. This is interesting that they seemed to be able to find different sets of interacting SNPs.

There is no general rule for selecting the best feature selection method in machine learning studies. The decision mostly depends on the data and research question of the investigation. For the purpose of detecting gene-gene interactions, Relief-based methods were shown to have better performance than the common univariate methods. Gene-gene interactions can be very challenging to detect by univariate methods since individual genetic factors may not show significant main effects. By comparing samples using all genetic attributes, Relief-base algorithms are able to capture the non-addition interaction effects among multiple attributes, and are recommended for detecting gene-gene interactions for GWAS.

Chapter 4

Ensemble Learning for Biomarker Discovery

4.1 Background

After reducing the size of the dataset using FS methods, two ensemble learning algorithms are used to model the genetic behaviors in the reduced genetic dataset. ML methods are applied to the dataset to detect either interactions among SNPs or identify variables with high marginal effects. ML methods are known to be the most widely used approaches in GWAS. Many successful studies in GWAS reported significance of ML methods in revealing the causal genetic variants in disease datasets [81]. ML methods are capable of regression and classification that are useful when there are quantitative and categorical phenotypes. In addition, these methods are promising complements to standard single-SNP tests and appropriate alternatives for multi-SNP analyses [81]. For example, nonparametric approaches such as ensemble methods can

be used to model complex interacting relationships among multiple SNPs.

4.1.1 Machine Learning Methods

The parametric linear statistical models have limitations for detecting non-linear patterns of interactions [58]. Likewise, most of the ML methods are single-variable approaches in which interactions between multiple variables are not investigated, rather, the main effect of one variable is considered [59]. It is discussed by Moore et al. that the data mining and machine learning methods can reveal numerous significant interactions and other complex genotype–phenotype relationships when they are widely applied to GWAS data [58]. In addition, these computational approaches make fewer assumptions about the data and functional form of the model.

There are other categories of ML algorithms such as: instance-based (e.g., KNN), rule system (e.g., Cubist), regularization (e.g., ridge regression), neural networks (e.g., back-propagation neural networks), and clustering (e.g., k-Means) which are not considered in this study. The reason is that these methods are not suitable for the CRC dataset, in spite of the ensemble methods. The instance-based and regularization methods are simple linear methods which seem inappropriate for genomic studies [81]. In addition, clustering algorithms are not helpful in this situation because in this study, the focus is classification.

4.1.2 Ensemble Methods

Ensemble methods use a set of predictors known as base learners. To produce a final prediction, the predictions of the base learners are weighted and the overall predic-

tions are decided as majority voting, i.e., the most voted class label, (for classification) or the average of fitted values (for regression). It has been shown that the ensemble methods perform better than other approaches under certain circumstances [21]. Firstly, the components of ensemble methods should be weak learners such as classification and regression trees (CART) [7]. Secondly, the base learners have to be different from each other, meaning a reasonable variance should exist among them. One popular method is bagging (short for bootstrap aggregating) that works based on bootstrapped samples of the original data [7]. Bagging is a technique for reducing the variance of an estimated prediction function. It seems to work especially well for high-variance, low-bias procedures, such as trees [27].

Random Forests (RF) are a special case of bagging in which more randomness is added such that the variables are randomly selected to determine the optimal split at each node of the tree [8]. The trees in RF are uncorrelated. The RF algorithm is an effective prediction tool which can uncover interactions among genes that do not exhibit strong main effects [58]. RF have been utilized in various studies including: to predict rheumatoid arthritis status using SNPs [78], to rank SNP predictors [74, 77], and to identify the epistatic effects related to human diseases [29].

Another method to generate an ensemble is boosting in which, in contrast to bagging, the weak learners evolve over time and make weighted votes [27]. Here, each weak learner is weighted based on the result from the previous base learner. An example a boosting algorithm is the gradient boosting machine (GBM) which uses trees as base learners and minimizes the loss function using gradient descent [28]. The RF and GBM approaches provide variable importance measures that can be used to select the most relevant predictors [8, 28, 81].

4.1.3 Previous Works

Szymczak et al. reviewed applications of several ML methods (penalized regression, ensemble, and network-based) on three GWAS datasets [81]. The Ridge regression method was used to detect SNPs associated with the rheumatoid arthritis (RA) phenotype and resulted in identifying an SNP near the *HLA*-B gene [79]. D'Angelo et al. combined the least absolute shrinkage and selection operator (LASSO) with the principal component analysis (PCA) which detected two significant gene-gene interactions in RA data [18]. These studies showed that the penalized regression methods are not computationally feasible to be applied to GWA data simultaneously, rather, they require improved implementations or a reduced size of the data [81].

Further on, the RF ensemble method was applied to the RA data and identified many known and several new SNPs contributing to the phenotype risk [82, 88]. In another study, it was shown that RF can better identify and rank the causal SNPs and important interacting covariates when the Gini index (GI) variable importance measure is used for evaluating feature significance [42]. Bayesian network analysis (BNT) is a network-based approach which is used to detect relationships between predictor variables and the binary coronary artery calcification (CAC) phenotype in a simulated dataset. The results showed that only some of the known relationships were recovered in the BNT analysis [95].

Goldstein et al. recommended that using RF with default settings of hyper-parameters would not yield appropriate results for large GWA datasets. In contrast, tuning of different values of the RF hyper-parameters, *mtry* or random number of variables, and *ntree* or number of trees, specifically using higher values, work well for

large GWAS datasets [31]. In a recent study, Wright et al. investigated ability of RF in detecting gene-gene interactions in a simulated data [92]. In their extensive simulations, many factors such as interaction models, varying marginal- and interaction-effect sizes, minor allele frequencies (MAF) and *mtry* were considered when creating a simulated data. Two single and three pairwise variable importance measures were investigated on five interaction models. Results of the simulations showed that single variable importance measures could capture the main effects but failed in detecting interactions. The results of pairwise variable importance measures indicated that they cannot detect the interaction in the presence of marginal effects. With all measures, marginal effects were detected as interaction effects and true interactions were not found. The reason for all of these is that current variable importance measures in RF cannot differentiate between marginal and interaction effects.

A data driven study by Olsen et al. revealed that tree-based ensemble ML methods outperform other methods such as Support Vector Machine (SVM) and Naive Bayes methods in the classifications of bioinformatics data [63]. Thirteen ML methods were compared on 165 bioinformatics datasets based on their performance producing higher classification accuracy, i.e., 10-fold cross-validation (CV) accuracy. The comparison of these ML methods on 165 different datasets demonstrated that Gradient Tree Boosting and subsequently RF outperformed other methods in terms of performing classifications in bioinformatics data.

4.2 Methods

Overall, due to their intrinsic multivariate and non-linear properties, tree-based ensemble methods prove to be a powerful analysis tool in the context of GWAS. In terms of risk prediction, tree-based methods are shown to be very effective in classifying individuals given their genotypes, while in terms of loci identification they are confirmed to be a well-suited alternative to standard approaches [58, 83].

An advantage of the RF approach is that the final decision trees may reveal interactions among SNPs that do not exhibit strong main effects [15, 16, 58]. The RF method is a non-parametric approach and is able to model the non-linear relationships among attributes [100]. RF are robust in the presence of noisy or potential false positive SNPs [10]. The primary limitation of tree-based methods is that they take marginal effects of variables into account. That is, the RF algorithm finds the best single variable for the root node before adding additional variables as nodes in the model.

GBM is a statistical learning method that can capture SNP-SNP and SNP-covariate interactions. In each split of a tree, all variables are considered jointly for associations with the phenotype and the variable that best increases classification accuracy is selected for that split. Depending on the depth of the tree in GBM method, the higher order interactions can be detected by the model. In addition, no specific genetic model (e.g., additive, dominant, recessive) is specified a priori, rather GBM models are built in a data-driven basis. It has a much lower computational burden compared to RF and even performed as well or better than RF in a study [34, 51].

Important note to consider when performing ML modeling is the determination of the significance of a variable. Significance of a variable in a model is implied if the inclusion of that variable in a given model leads to better modeling (i.e., higher prediction accuracy) of the given dependent variable, compared to when it is left out. Depending on the algorithm being used, some of them detect interactions and some of them only quantify the significance of variables in a dataset. Therefore, a variable is considered significant if its inclusion improve the performance (e.g., prediction accuracy) of the model.

The essential part of the RF and GBM ensemble methods is parameter tuning. We perform different runs of these algorithms with different parameter values. The dataset is randomly partitioned into 10 folds. Each time an ML model is trained on 9 folds and evaluated on the other fold. This is repeated for all 10 folds, resulting in 10 different models. The accuracy is then calculated as the averaged accuracy of these 10 models. Since randomness may affect the results of predictions, we repeat each round (of 10-fold CV) 10 times. That is, each ML model (with a specific parameter settings) is repeated 100 times and the accuracy is averaged over these 100 times. In addition, since an SNP has categorical values {0,1,2}, we convert these values to factors so that ML methods treat them as categorical values.

The most significant SNPs of the ML methods are compared to determine if they have intersections. If these ML methods give similar top variables we can state that those variables are of great importance, or that maybe they are the causal factors of the disease. For validation, the results are compared to the findings from other studies or online databases which describe biological characteristics of CRC. The algorithm and implementation details of these two ML methods are explained as follows.

4.2.1 Random Forests

Random Forests (RF) are shown to be a very powerful regression and classification method which are created from a large collection of possibly uncorrelated decision trees [8, 81, 100]. Each tree is grown using the CART methodology [7]. The most significant feature in the RF is that for the kth tree, a random vector Θ_k is generated, independent of the past random vectors $\Theta_1, ..., \Theta_{k-1}$ but with the same distribution. If the input vector in the original data has M variables, then Θ_k would have a randomly selected number of $m < M$ variables. These m variables are used to make the best split at each node of the tree. The tree is then grown using a training set $\mathbf{x}$ (which is sampled at random with replacement) and Θ_k resulting in a classifier $h(\mathbf{x}, \Theta_k)$. Each tree is grown to maximum size without pruning. Intuitively, reducing m will reduce the correlation between any pair of trees in the ensemble. For classification, the default value of m is $\lfloor\sqrt{M}\rfloor$; however, the best value for this parameter will depend on the problem, and it should be treated as a tuning parameter [8, 27].

Figure 4.1 shows the procedures of creating a decision tree in RF. RF uses bootstrapping to grow trees. Using the bootstrapping technique, usually one third of the training set is not present in growing trees. This left over data is known as the out-of-bag (OOB) data. The OOB data, which are not present in the training set, are replaced with duplicate samples to rectify the size of the data. After the development of trees, the OOB samples are used to test the individual trees as well as the entire forest. The average misclassification error over all trees is known as the OOB error estimate. Accuracy of RF depends on the strength of the individual tree classifiers and also the lack of correlation between trees.

After a large number of trees are generated, RF vote for the most popular class as the result of classification. That is, after creating all trees, the new entry goes down from all of the trees to obtain a class (case or control) vote for each tree. From the result of the classifications, the class with the highest vote (among all trees) is considered as the prediction for that entry.

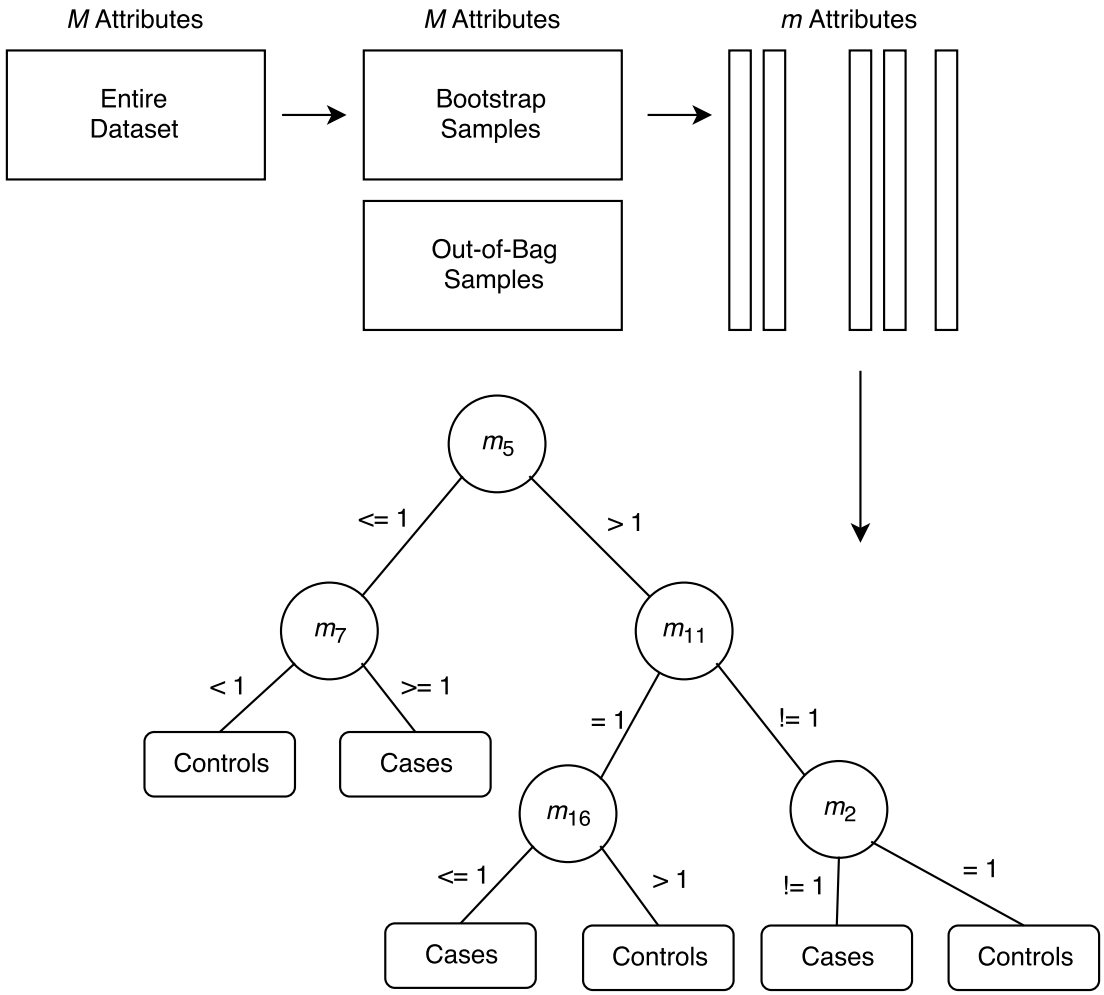

Figure 4.1: Overview of the RF algorithm, adopted from [67].

Another important feature of the RF algorithm is the variable importance cal-

culation. This algorithm analyzes each attribute and reveals the importance of the attribute in predicting the correct classification in each tree. RF gives estimates of variables' significance in the classification using the permutation or Gini importance measures. In the permutation importance, the values of the jth variable are randomly permuted in the OOB samples, and then samples are reclassified using these new values. The number of correct classifications with the permuted values is compared with the number of correct classifications in the original data. The decrease in accuracy as a result of this permuting is averaged over all trees, and is used as a measure of the importance of variable j in the random forest [27]. That is, if randomly permuting values of variable j does not affect the predictive ability of trees on out-of-bag samples, that attribute is assigned a low importance score [10]. The drawback of permutation importance is the computational burden when the number of variables is huge, which is the case for GWAS data.

In the calculation of the Gini importance, in every split of a node on the jth variable, the Gini impurity criterion over all trees is calculated as denoting the importance measure of that variable. In this measure, the Gini index from a single tree is generalized to a forest [100]. Let p_k be the proportion of observations of outcome class k at a node. The Gini index of node S_k is a measure of impurity i and is then given by $i(S_m) = 1 - \sum_k {p_k}^2 = \sum_{k;l;k\neq l} p_k p_l$. The impurity of a tree t, i.e., the Gini index is the sum over all terminal nodes S_m of the impurity of a node $i(S_m)$ multiplied by the proportion p_m of subjects that reach that node of the tree $\text{Gini}(t) = \sum_m p_m.i(S_m)$. This measure is extended for all trees, which more explanation can be acced in [8, 100].

We use a very fast implementation of RF provided in an R package named

'ranger' [91]. The 'ranger' package provides all functionalities similar to the 'RandomForest' package in R with much faster implementation speed. Therefore, we can use it for the GWAS datasets with a large number of SNPs. As for Brieman's random forests, 'ranger' accepts two main parameters: *ntree*, the number of trees; and *mtry*, the number of random features at each node. Expert knowledge was used to choose ranges of values for these parameters. For *mtry* these values were selected as: $mtry = \{100, 200, 300, 500, 1000, 2000\}$. Values for *ntree* are: $ntree = \{500, 1000, 2000\}$. Other parameter values remained as package default. There are 18 different configurations of *mtry* and *ntree* in which each combination is repeated 10 times. That is, RF is run 180 times (or 1800 times when 10 runs of cross-validation are included.)

4.2.2 Gradient Boosting Machine

At first, in a boosting algorithm, many weak learners are built and the new learners focus on improving the previous ones. Based on the definition by Hastie et. al [27], a weak classifier is one whose error rate is only slightly better than random guessing. The learners are trained sequentially, which result in building a "committee" of complex predictors [27, 28]. GBM is a boosting ML algorithm in which a weighted combination of predictors are used to make the final prediction [28].

Using numerous base learners, a set of approximations $F_m(x); m = 1, 2, ..., M$ is created. The "error" in the predictions is calculated based on a loss function $L(y, F(x))$ such as squared-error $SE = \sum (y - F)^2$. The F is then adjusted to $F(x) = F(x) + \rho \times h$ in which ρ is a regularization parameter (or coefficient) and

h is base learner with parameters optimized from the gradient of the loss function $\nabla L(y, F)$.

The listing 2 shows the description of the GBM algorithm [28]. In this algorithm, M is the number of base learners (i.e., trees), N is the number of training samples, F is the approximation function, L is the loss function, $h(\mathbf{x}; \mathbf{a})$ is the base learner that fits training data $\mathbf{x}$ with a set of parameters $\mathbf{a} = \{a_1, a_2, ...\}$, ρ is the set of coefficients for base learners, and β is coefficient for base learners when optimizing parameters $\mathbf{a}$.

Algorithm 2 Gradient boosting algorithm

1: **procedure** GRADIENT_BOOST

2: $F_0(\mathbf{x}) = argmin_\rho \sum_{i=1}^{N} L(y_i, \rho)$

3: **for** $m = 1$ to M **do**:

4: $\tilde{y_i} = -[\frac{\partial L(y_i, F(\mathbf{x}_i))}{\partial F(x_i)}]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}$, $i = 1, N$

5: $\mathbf{a}_m = argmin_{a,\beta} \sum_{i=1}^{N} [\tilde{y_i} - \beta h(\mathbf{x}_i; \mathbf{a}_m)]^2$

6: $\rho_m = argmin_\rho \sum_{i=1}^{N} L(y_i, F_{m-1}(\mathbf{x}_i) + \rho h(\mathbf{x}_i; \mathbf{a}_m))$

7: $F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \rho_m h(\mathbf{x}_i; \mathbf{a}_m)$

8: *endFor*

9: *end Algorithm*

At line 2, an initial guess $F_0(\mathbf{x})$ is made on the training data, and then from line 3 the models for M iterations are built. Starting from line 4, for each step m, first the negative gradient of loss function $\tilde{y_i}$ over all training samples is calculated. Second, the optimal parameters $\mathbf{a}_m$ are found such that the least-squares of $\tilde{y_i}$ and base learner is minimized. Third, given the $h(\mathbf{x}; \mathbf{a})$, the optimal value of the coefficient

ρ_m for model m is determined. Last, the approximation $F_m(x)$ is updated as shown at line 7. This procedure is repeated for all learners until the final prediction F_M is obtained.

GBM can also be used to rank-order SNPs according to their cumulative predictive performance. The variable importance measure used in GBM is similar to the Gini importance commonly used in Random Forests [8, 28]. Therefore, the measure of importance can be used to identify significant SNPs in the GWAS dataset.

We use an R package called 'gbm' for performing classifications based on GBM [68]. Similar to the RF, the range of values for hyper-parameters are chosen based on expert knowledge. GBM has three main parameters: *n.trees*, the number of trees; *interaction.depth*, the complexity of interactions between nodes (i.e. features); and *shrinkage*, the learning rate or step-size reduction in the GBM algorithm. The values of these parameters are as follows: $n.trees = \{100, 500, 1000, 2000\}$, $interaction.depth = \{1, 2, 10\}$, and $shrinkage = \{0.001, 0.01, 0.1\}$. There are 36 different configurations of these three parameters in which each combination is repeated 10 times. That is, GBM is run 360 times (or 3600 times when 10 runs of CV are included). Other parameters of GBM such as *n.minobsinnode*, *bag.fraction*, and *train.fraction* remained default to 10, 0.5, and 0.5 respectively.

4.3 Results

In this section, we explain the results of applying ML methods to the CRC dataset. The ensemble algorithms are applied to the reduced dataset and their classification accuracy for different parameter settings are recoded.

4.3.1 Applying Random Forests to CRC Dataset

We apply RF to the CRC dataset with a reduced feature set of 2,798 SNPs as a result of the TuRF feature selection algorithm. Figure 4.2 shows the average accuracy and the area under the ROC curve (AUC) for 18 runs of RF with different combinations of parameters. The x-axes are *ntree*, i.e., the number of trees, and the y-axes are the average accuracy and AUC. The different lines connect values of accuracy and AUC for different *mtry* values. The measures of accuracy and AUC for each configuration of *mtry* and *ntree* are averaged over 10 different runs. From this plot, we see that the highest accuracy of 75% and AUC of 0.84 are obtained when $mtry = 100$ and $ntree = 2000$. The best accuracy is obtained when *ntree* has the maximum value and *mtry* has minimum value. Increasing *ntree* consequently increases accuracy, while increasing *mtry* decreases the accuracy. That is, for RF, the higher values of *ntree* and lower values of *mtry* are preferable in the context of GWAS data (at least for the CRC data).

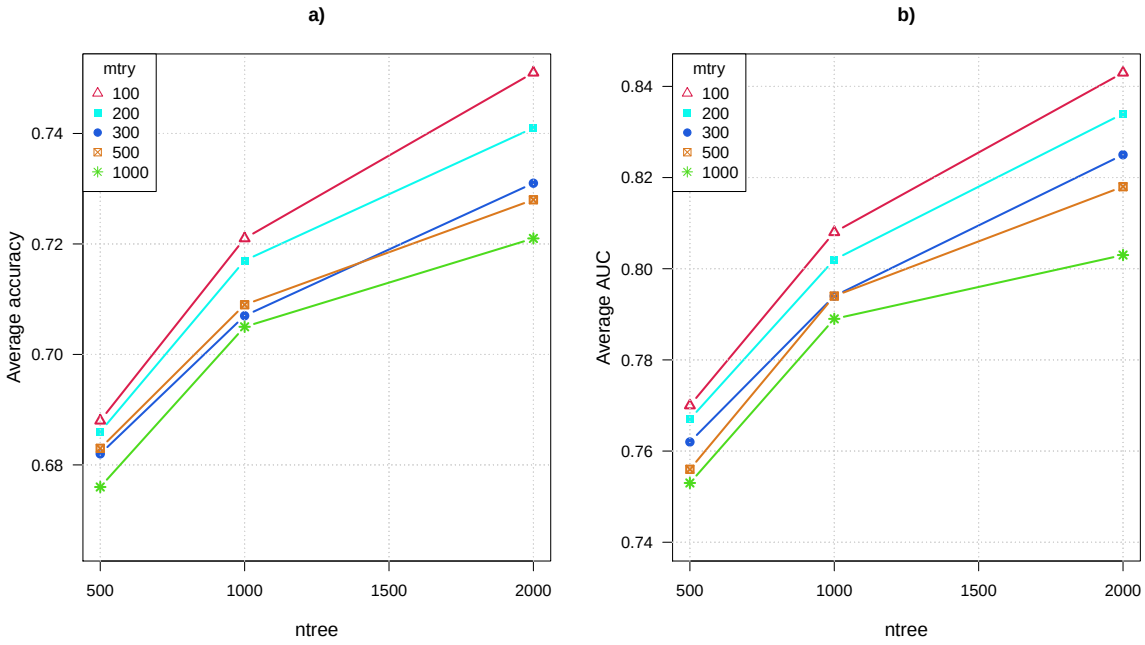

Figure 4.2: Parameter comparison for RF. **a)** shows the average accuracy of RF for different parameter values. **b)** shows the average AUC of RF for those parameter values.

The RF method would also give importance score to features based on their significance of effect on the class labels. In the 'ranger' package we choose the 'impurity' as the measure of importance score. Therefore, SNPs are assigned an importance score (of between 0 and 1) for a run of the RF model. We run the RF method 10 times with the hyper-parameters $mtry = 100$ and $ntree = 2000$. The SNPs' importance score are added up for all 10 runs of the RF model (or 100 runs when the runs of 10-fold CV are included). At the end, each SNP would be assigned an average importance score which is the significance of that SNP over all runs of the RF implementation. The distribution of the average significance score of SNPs is shown in Figure 4.3. From

the plot, we see that a high proportion of SNPs have a low average score, and only a small proportion of SNPs have a high average score.

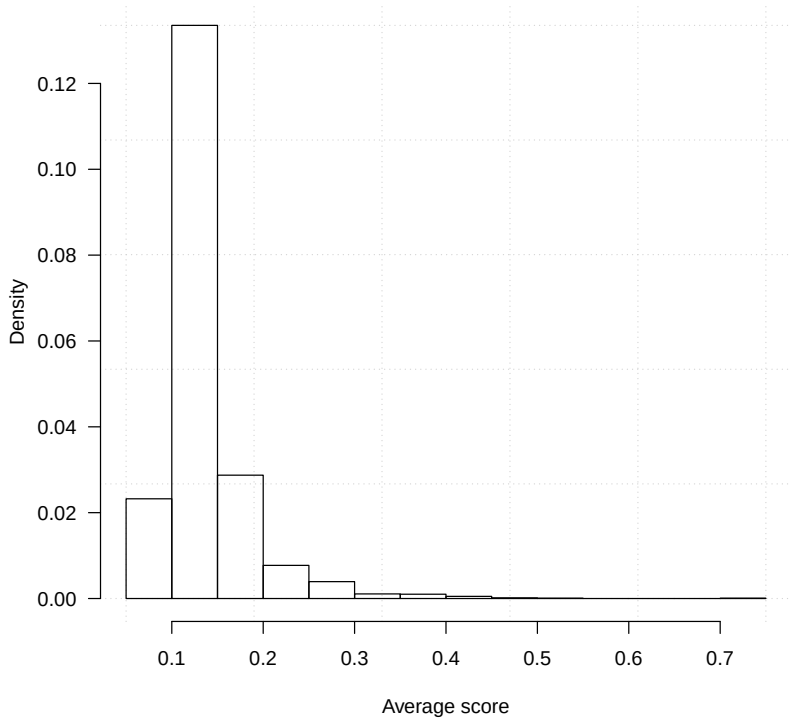

Figure 4.3: Plot of SNPs' average score by RF

4.3.2 Applying Gradient Boosting Machine to CRC Dataset

We then apply GBM to the CRC dataset with selected 2,798 SNPs. Figure 4.4 shows the average accuracy and the area under the ROC curve (AUC) for 36 runs of the GBM with different combinations of parameters. In this figure, the x-axes of subplots show the *n.trees*, i.e., the number of trees, and y-axes of subplots show the average accuracy and AUC for different *interaction.depth* values. Each row in

this plot shows the average accuracy and AUC for a value *shrinkage* parameter while *interaction.depth* values differ.

From this figure, we see that the highest average accuracy is 74% and AUC is 0.82, which are obtained when $n.trees = 2000$, $interaction.depth = 10$, and $shrinkage = 0.1$. GBM performs weakly with lower values of *n.trees* while increasing the iterations results in better predictions. As the number of iterations (i.e., number of trees) increases, the accuracy gets higher. In addition, the $interaction.depth = 10$ has the highest accuracy, meaning that the more complex interactions among SNPs result in better performance by the GBM. For all of the GBM models, as *interaction.depth* increases, the accuracy gets increased. Similar to the *interaction.depth*, increasing *shrinkage* also improves the accuracy of GBM models. That is, *shrinkage* of 0.1 gives better predictive accuracy than lower values such as 0.01 and 0.001. For GBM, the higher values of hyper-parameters of *n.trees*, *interaction.depth*, and *shrinkage* are preferable in the context of GWAS data (at least for the CRC data).

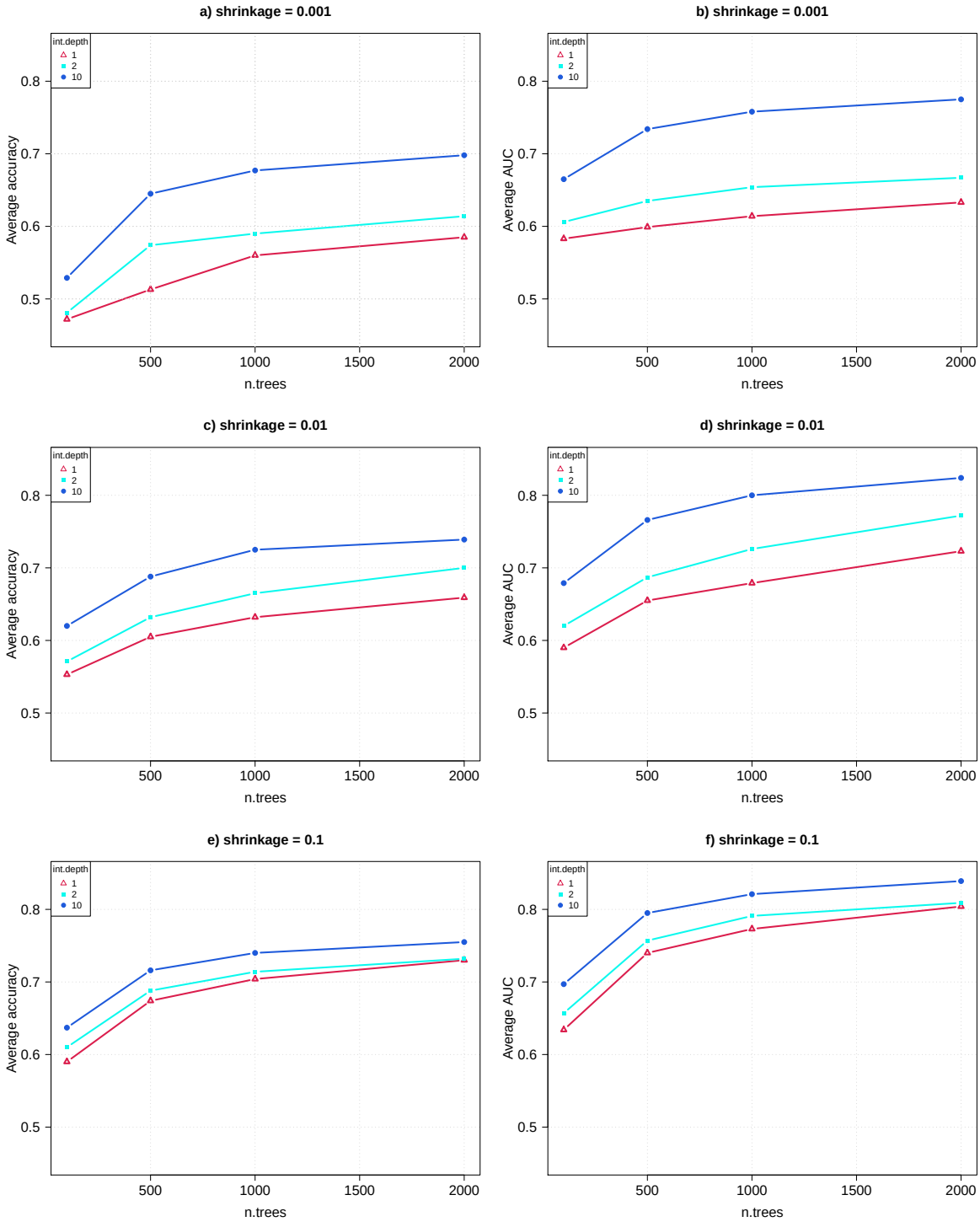

Figure 4.4: Parameter comparison for GBM. **a) c) e)** show the average accuracy for *shrinkage* of 0.001, 0.01, 0.1. **b) d) f)** show the average AUC for *shrinkage* of 0.001, 0.01, 0.1 respectively.

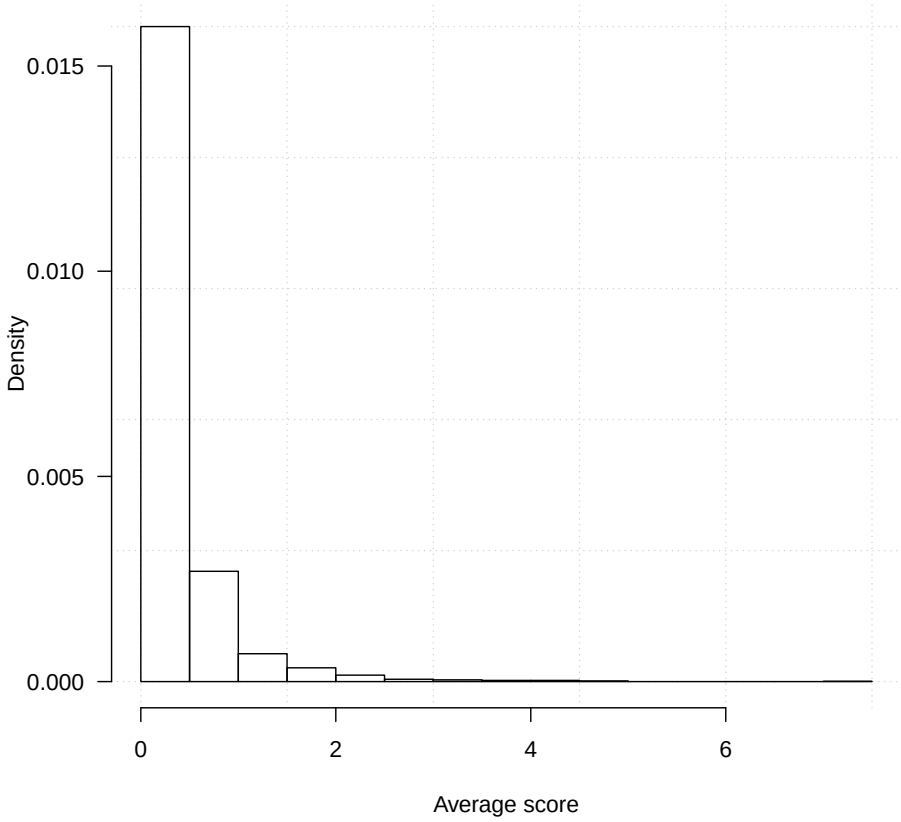

Figure 4.5: Plot of SNPs' average score by GBM

Similar to RF, GBM also gives an importance score to features based on their significance of effect on the class labels. SNPs which have high effects on the class label, would be assigned a high score. For a run of the GBM model, SNPs are given an importance score that in contrast to RF can be more than 1. We use the best configuration of GBM hyper-parameters to detect the most important SNPs in the dataset. We repeat GBM with hyper-parameters of the best model, with hyper-parameter values $n.trees = 2000$, $interaction.depth = 10$, and $shrinkage = 0.1$, for

10 times and record the significance score of every SNP. The SNPs' importance score are added up for all 10 runs of the RF model (or 100 runs when the runs of 10-fold CV are included). At the end, each SNP would be assigned an average importance score which is the significance of that SNP over all runs of the GBM model. The distribution of the average significance score of SNPs are shown in Figure 4.5. From the plot, we see that a high proportion of SNPs have a low average score.

4.3.3 Key Genetic Markers Discovered by RF and GBM

We applied two ensemble algorithms to the reduced CRC dataset with 2,798 SNPs to identify the most significant variants associated with the disease phenotype. Consequently, GBM and RF methods produced significance score for SNPs. After obtaining scores for SNPs by the ensemble methods, we compare the most important SNPs by these methods and choose the ones which are considered as important by both of them. Figure 4.6 shows the plot of SNPs in which the x-axis is the average importance score by GBM and the y-axis is the average significance score by RF. Therefore, the SNPs which are in the top-right corner of the plot are found to be important by both methods. These SNPs have a high significance score by both RF and GBM methods. We see that there are almost two clusters of SNPs. The first cluster is in the bottom-left corner of the plot where most of them have very similar scores. The other cluster is in the top-right part of the plot, which could be grouped together. Therefore, a separating line is drawn to separate SNPs into two clusters. The SNPs on the right side of the line are the most significant SNPs which have the highest score by both RF and GBM methods. Moreover, SNP **rs3760948_T** is the Pareto-Front

of this plot because it has the highest score in both RF and GBM. At the end, based on the specified threshold, the 44 most significant SNPs which are detected by both methods are selected. Table 4.1 shows these 44 SNPs and their average scores by the RF and GBM methods.

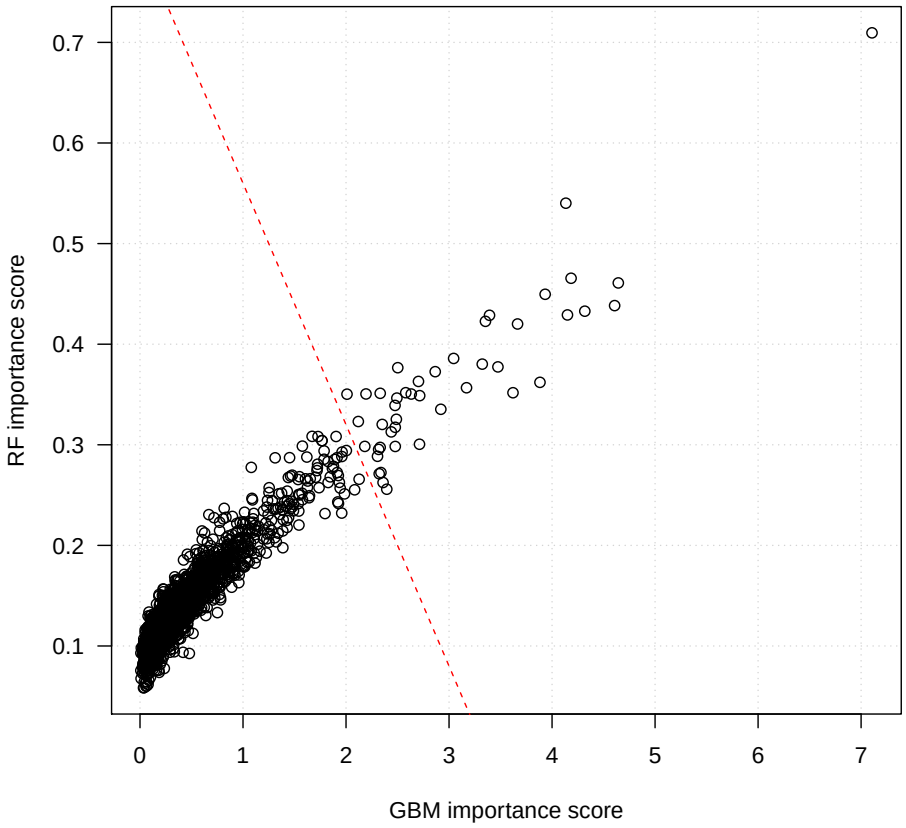

Figure 4.6: Scatter plot of SNP scores by the two ensemble learning algorithms. The x-axis shows the GBM importance scores and the y-axis shows the RF importance scores.

Table 4.1: The 44 most important SNPs from the ensemble learning algorithms

Name	RF_score	GBM_score	Name	RF_score	GBM_score
rs3760948_T	0.7095	7.1059	rs4625115_T	0.3505	2.1938
rs7594717_G	0.5402	4.1339	rs3844138_A	0.3505	2.6331
rs12407198_G	0.4656	4.1847	rs17379465_A	0.3503	2.0094
rs9688110_A	0.4609	4.6431	rs1212694_A	0.3489	2.715
rs2571219_G	0.4496	3.933	rs10016091_G	0.3464	2.4935
rs2386946_A	0.4384	4.6075	rs1991915_T	0.3392	2.4759
rs344570_T	0.4328	4.318	rs13263313_T	0.3353	2.9202
rs8022574_A	0.4291	4.1497	rs11610311_C	0.3253	2.4885
rs10814848_G	0.4287	3.3928	rs6578849_G	0.3232	2.1192
rs2179321_T	0.4227	3.3535	rs17831158_A	0.3202	2.3514
rs3912454_C	0.4201	3.6651	rs2010907_G	0.3174	2.4778
rs6782709_G	0.3858	3.0446	rs11783793_T	0.3128	2.4392
rs898438_G	0.3802	3.3224	rs17162736_A	0.3005	2.7131
rs2645737_C	0.3774	3.4745	rs11185516_A	0.2984	2.182
rs647831_G	0.3766	2.5022	rs11985944_T	0.2983	2.4779
rs658836_C	0.3727	2.8659	rs1816647_T	0.2974	2.3314
rs7747931_A	0.363	2.7031	rs2736486_C	0.2954	2.3167
rs4961513_A	0.3621	3.8827	rs721619_G	0.2886	2.3056
rs1495008_C	0.3567	3.1697	rs12695485_T	0.2724	2.3389
rs2406370_G	0.3518	2.5797	rs952880_C	0.2712	2.3221
rs9288684_T	0.3517	3.6207	rs1367128_G	0.2625	2.3595
rs1505229_T	0.3512	2.3319	rs3842986_T	0.2558	2.3955

4.4 Biological Interpretation

Finding the genetic risk factors in the CRC dataset is just one step toward revealing the etiology of the disease. The next step after finding the most significant genetic variants via ML methods in GWAS is to conduct a biological validation on the findings using online databases. Online biological databases contain genetic information about SNPs and describe the functions of the corresponding gene regarding a genetic variant in DNA sequence. By exploring these sources, more information can be acquired about SNPs and the association of the genes with disease phenotypes.

Table 4.2 shows more information about the most important SNPs found by the ensemble algorithms. In this table, CHR represents the chromosome number; SNP is the id of SNP; A1 is the minor allele of SNP; MAF is the frequency of minor allele; and P-value is the p-value of the association of the SNP with the disease (which is obtained with PLINK). We see that most of the SNPs have a very low p-value indicating the significance of their association with the disease. We explore the ENSEMBL[1] and National Center for Biotechnology Information (NCBI)[2] databases to find the genes associated with these significant SNPs. The corresponding genes of these SNPs are shown in the seventh column of Table 4.2. We see from Table 4.2 that 15 (out of 44) SNPs are in non-coding regions and the remaining 29 SNPs, which most of them have *intron* functional class, are in the protein coding regions of DNA. These 29 genes are of great importance, because in the pathway they are transcribed to RNA and result in producing proteins may lead to causing CRC.

[1] http://www.ensembl.org

[2] https://www.ncbi.nlm.nih.gov/

Table 4.2: List of the 44 identified SNP markers

CHR	SNP	A1	MAF	P-value	Gene
1	rs647831	G	0.345	0.01041	-
1	rs12407198	G	0.3376	0.005675	C1orf101
2	rs1816647	T	0.3906	0.03521	-
2	rs7594717	G	0.3347	**4.632e-05**	ALK
2	rs1505229	T	0.3912	0.0009559	LRRTM4
2	rs1367128	G	0.1925	0.003527	THSD7B
2	rs9288684	T	0.07827	0.0001041	INPP5D
3	rs12695485	T	0.111	0.02966	LOC107986044
3	rs6782709	G	0.345	0.01885	LOC105374217
3	rs11185516	A	0.4273	0.6691	ZDHHC19
4	rs1991915	T	0.3658	0.006061	OTOP1
4	rs10016091	G	0.4641	0.002475	SCFD2
4	rs2736486	C	0.3259	0.01471	-
4	rs2010907	G	0.2764	0.0008263	-
5	rs2406370	G	0.4361	0.163	ITGA1
5	rs9688110	A	0.3562	0.0009197	FAT2
6	rs7747931	A	0.4329	0.04637	E2F3
6	rs952880	C	0.4848	0.2318	KCNQ5
7	rs17379465	A	0.3155	0.1253	-
7	rs17162736	A	0.1398	0.01466	STEAP2-AS1
8	rs11985944	T	0.2692	0.001981	-

8	rs11783793	T	0.4169	0.0006035	-
8	rs721619	G	0.3352	0.1755	EPHX2
8	rs17831158	A	0.3896	0.001942	LINC00968
8	rs1495008	C	0.1701	0.00374	LOC101929628
8	rs13263313	T	0.3472	0.008509	JPH1
9	rs10814848	G	0.5072	0.001	GLIS3
9	rs3912454	C	0.4712	0.03876	-
9	rs4961513	A	0.2907	0.0003435	-
9	rs4625115	T	0.4018	0.0006728	-
11	rs6578849	G	0.3794	0.000583	SYT9
12	rs11610311	C	0.2652	0.002581	-
14	rs8022574	A	0.4058	0.002786	-
14	rs2645737	C	0.4497	0.01135	NID2
14	rs1212694	A	0.234	0.00103	ACTR10
18	rs658836	C	0.2109	0.0004416	-
18	rs898438	G	0.3658	**0.0009927**	DCC
18	rs2571219	G	0.3866	0.001339	ATP8B1
18	rs3844138	A	0.2504	0.0112	-
19	rs3760948	T	0.3714	0.0002007	ARRDC5
19	rs344570	T	0.08866	0.0002718	TNFSF14
20	rs2179321	T	0.5128	0.04872	PLCB4
20	rs2386946	A	0.2101	0.005192	CDH4
21	rs3842986	T	0.2244	0.04063	-

Figure 4.7 shows the distribution of SNPs and the number of genes in the chromosomes. In this plot, the coding and non-coding genes in all of the chromosomes are shown. We see that there are no SNPs in the chromosomes {10,13,15,16,17,22} and the chromosome 8 contains 6 SNPs in which 4 of them belong to coding regions. Similarly, for chromosome 2, 4 out of 5 SNPs belong to the coding regions of DNA.

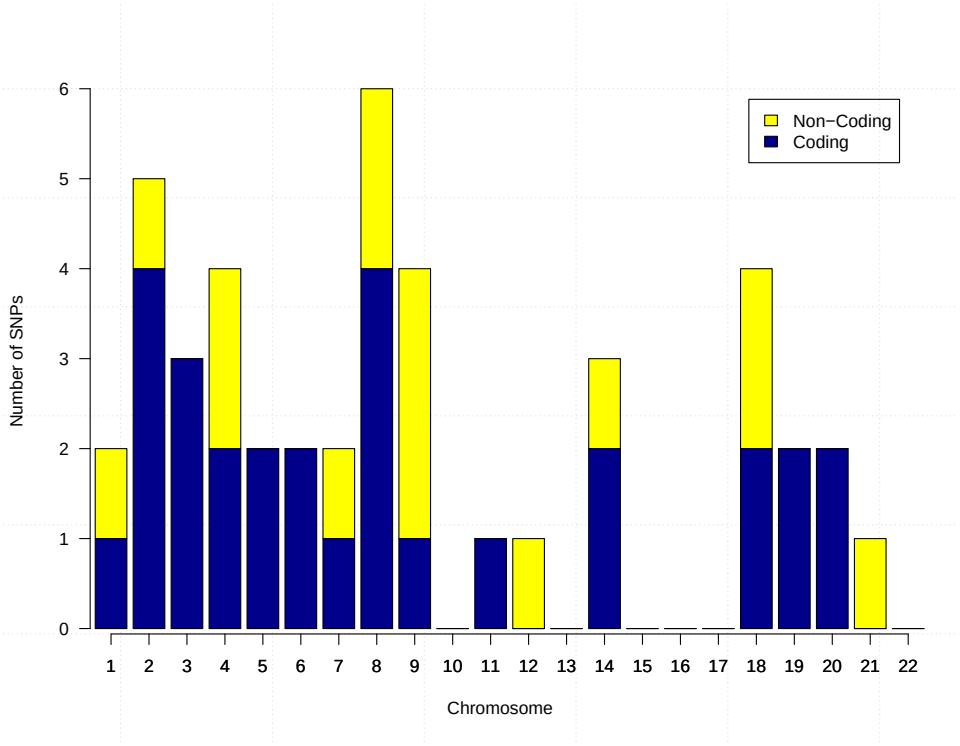

Figure 4.7: Number of coding and non-coding SNPs in each chromosome

The homogeneity and heterogeneity of SNPs can reveal important information. The distribution of SNPs' value among the 472 cases indicates that most of these SNPs are heterogeneous as shown in the Figure 4.8. That is, they have values of 1 and 2. The x-axis is the SNP name and the y-axis is the number of cases. Each bar in this plot shows the distribution of SNP values {1,2} among cases. From the

figure, we see that SNPs have more value of 1 than 2, meaning that they are mostly heterogeneous. This heterogeneity is important since it may be interpreted as the primary cause of the disease.

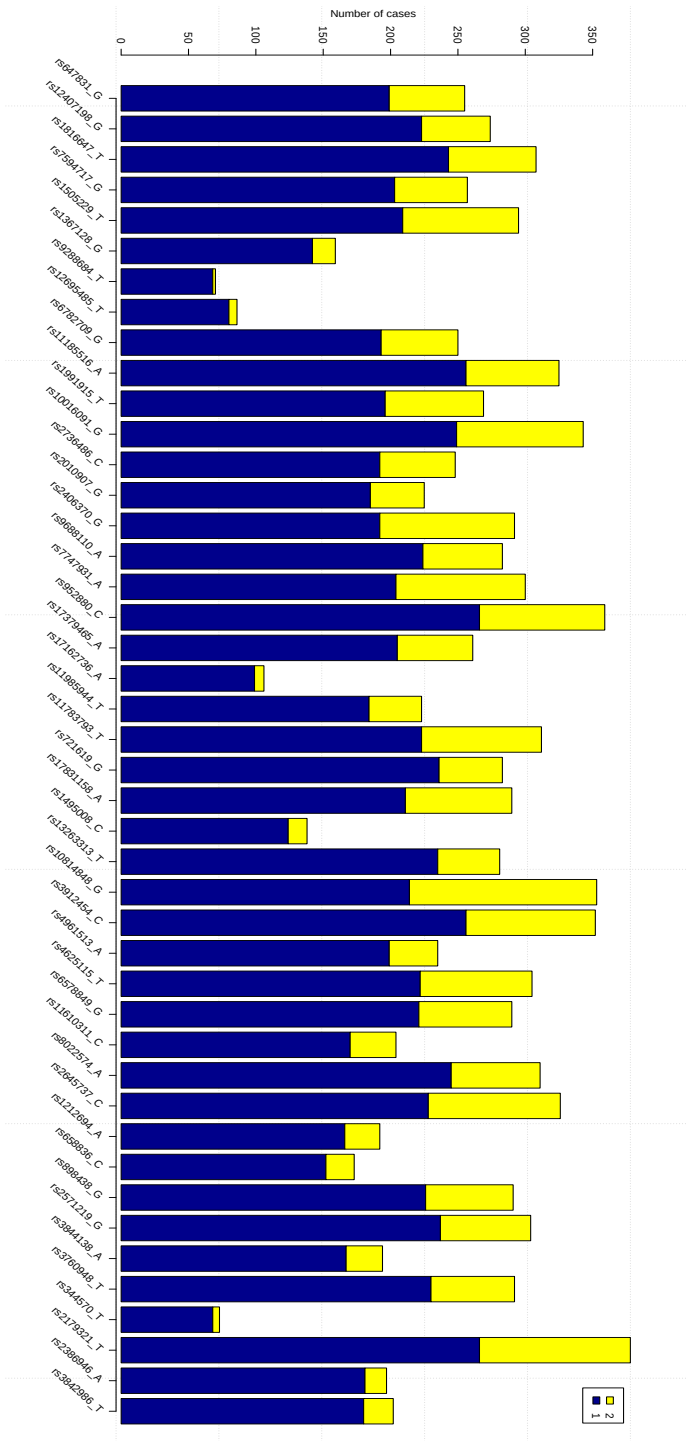

Figure 4.8: SNPs heterogeneity and homogeneity among cases

4.4.1 Detailed Information on the Genes

Even though expert knowledge is needed to investigate the functions of the above genes, their existing associations with the disease can be determined through biological databases. Therefore, we explore detail information of these genes using the ENSEMBL database which provides information about genes and their associated diseases.

Among the above 29 genes, there are a few genes which are known to be associated with the disease phenotype including: DCC, ALK, ITGA1, E2F3, and NID2. The most important gene is **DCC** which is known to be directly associated with *colorectal cancer* based on the ENSEMBL database[3]. The gene expression, function, biological features, and molecular genetics of DCC show associations with colorectal cancer[4]. It was shown by Castets et al. that DCC functions as a tumor suppressor in the colorectal cancer [12].

Gene **ALK** is directly related to the phenotype *colorectal adenocarcinoma sample* based on a number of studies. Pietrantonio et al. discussed that gene ALK may prevents the effects of other treatments for advanced colorectal cancer. Fusions of genes ROS1 and ALK occur in colorectal cancer and may have substantial impact in the treatment of the disease [1]. In addition, Lipson et al. investigated 145 genes related to colorectal cancer in 40 tissues. They identified a gene fusion of ALK and another gene from colorectal samples which have major therapeutic relevance [50]. There is also a study performed in a Japanese population showing the association of ALK with schizophrenia [46]. Findings from a study by Slambrouck et al. showed

[3]http://www.ensembl.org/Homos_sapiens/Gene/Phenotype?db=core;g=ENSG00000187323

[4]https://www.omim.org/entry/120470

that the α1-integrin (i.e., **ITGA1**) is relevant to the prevention of tumor progression in colon cancer patients [86]. In addition, Boudjadi et al. analyzed the tissue microarrays from 65 patients revealing a clear correlation of ITGA1 expression in 72% of the colorectal cancer patients [6]. Akao et al. showed that **E2F3** is involved in the prevention of colorectal cancer [2]. A GWAS study on a Chinese population shows that variants of the gene **NID2** are associated with the phenotype *lung cancer (smoking interaction)* [99].

4.4.2 Enrichment Analysis

After finding genes related to the disease, it is important to identify pathways and genetic ontologies (GO) leading to the causes of the disease. There are numerous enrichment analysis tools which are differently categorized based on their background algorithms, different enrichment analysis methods, and correction testings. Some of these tools include: Onto-express, FunSpec, GeneMerge, MAPPFinder, GoMiner, GARBAN, FuncAssociate, EASE, and DAVID [40]. In this study we use the Database for Annotation, Visualization and Integrated Discovery (DAVID) [20], because it is widely used for enrichment analysis, specifically for exploring the functions of genes. It is a web-based tool which takes a list of genes as an input and produces many annotation tables and charts, based on the functional annotation clustering, gene functional classification, and gene name batch viewing algorithms. These algorithms are useful for identifying the disease and relevant GO terms associated with a given list of genes.

We feed the above 29 genes to DAVID and it converts them to 3,873 DAVID

IDs which are the other names for these genes (obtained from different databases) or the genes known to be related to them. Then, we use these 3,873 genes for further analyses by DAVID. From the DAVID user interface, the annotations of Disease, Gene Ontology, and Pathway are selected for performing combined functional analysis. From disease category all disease databases are selected: *GAD_DISEASE*, *GAD_DISEASE_CLASS*, and *OMIM_DISEASE*. Similarly, for pathway annotation all databases are selected and from the gene ontology category six databases: *GOTERM_BP_DIRECT*, *GOTERM_BP_FAT*, *GOTERM_CC_*

DIRECT, *GOTERM_CC_FAT*, *GOTERM_MF_DIRECT*, and *GOTERM_MF_FAT*.

The functional annotation chart in DAVID shows all of the GO Terms related to the given list of genes based on the selected annotation categories and databases (which we selected as above). The threshold of 5 genes out of 29 genes is set to remove Terms which have less associated genes. The combined view for selected annotations are shown in Table 4.3. In this table, the column 'Category' shows the database; the column 'Term' is the GO term associated to the genes; 'Count' shows the number of genes (out of 29) included in that GO Term; and P-value indicates the permutation test of significance of the association.

All of the detected biological pathways have a p-value of less than 5%, even though most of them are general Terms that may apply for any disease. Nevertheless, there are 14 genes associated with the Term 'tobacco use disorder' with a p-value of 3.9×10^{-5}, which is highly significant. In a study, Watson et al. investigated the effect of tobacco use on increasing the risk of CRC using a retrospective cohort study of germline mutation, which showed that the tobacco use is significantly associated with increased risk of CRC [90].

Table 4.3: The functional annotation chart of the given gene list

Category	Term	Count	P-value
GAD_DISEASE	tobacco use disorder	14	3.9E-5
GAD_DISEASE_CLASS	chemdependency	14	3.0E-4
GAD_DISEASE_CLASS	metabolic	15	3.6E-3
GOTERM_MF_DIRECT	calcium ion binding	5	6.1E-3
GAD_DISEASE_CLASS	cardiovascular	13	6.5E-3
GOTERM_MF_FAT	calcium ion binding	5	6.7E-3
GOTERM_BP_FAT	movement of cell	7	1.2E-2
GOTERM_CC_DIRECT	integral component of membrane	12	1.7E-2
GOTERM_MF_FAT	metal ion binding	10	2.0E-2
GOTERM_BP_FAT	neuron development	5	2.2E-2
GOTERM_MF_FAT	cation binding	10	2.2E-2
GOTERM_BP_FAT	locomotion	6	2.5E-2
GOTERM_MF_FAT	ion binding	10	2.8E-2
GOTERM_CC_DIRECT	plasma membrane	10	3.1E-2
GOTERM_BP_FAT	cell migration	5	4.2E-2
GOTERM_BP_FAT	neuron differentiation	5	4.7E-2

Cell migration is another important GO term in Table 4.3, which is involved in association with cancers. When there are some tumor cells, some of them obtain the ability to get rid of tissue and immigrate; this is called metastasis. The tumor cells migrate and enter the blood and will go to other tissues and separate the cancer.

Metastatic cancer is cancer that has spread to other parts of the body. When colon or rectal cancer spreads, it most often spreads to the liver. Colorectal cancer happens when cells that are not normal grow in your colon or rectum. Therefore, 'cell migration' GO term has the potential of having association with CRC. Even though, further investigations accompanying with expert knowledge are required to study the revealed biological pathways and determine their associations with the CRC, to some extent, the detected pathways explain the relationship of those 29 genes with the disease.

4.4.3 Interaction Analysis

The ensemble learning methods detected 44 most significant SNPs and the biological interpretations revealed associations of these SNPs and their corresponding genes with CRC. Consecutively, we extend our exploration by analyzing pairwise and three-way interactions between those SNPs in association with the disease. The 44 most significant SNPs from the results of ensemble algorithms are used to investigate the pairwise interactions based on IG. We create the dataset of 44 SNPs for 944 samples and calculate the pairwise IG between all $\binom{44}{2}$ pairs of SNPs. We set the (optional) threshold of $p < 0.02$, i.e., 20 times out of 1000 permutations, to only keep interactions with significant p-value, which result in 17 pairs of interactions. Table 4.4 shows the SNPs, IG value, and p-value (which is the significance of IG comparing to 1000 times of permutation) for these 17 significant pairs.

From the Table 4.4, the maximum IG value is 1.3% for interaction between SNPs rs2010907 and rs3760948 with significance p-value of 0.002. There are six pairs of cod-

ing SNPs, i.e., SNPs with genes, having interactions in which of greatest importance is the interaction between SNPs in genes LOC107986044 from chromosome 3 and DCC from chromosome 18, mostly because DCC is previously known to be associated with CRC. In addition, gene INPP5D has three significant interactions with other genes such as PLCB4 and CDH4 that can be a sign of association of the gene INPP5D with CRC. However, no previous study has shown the association of INPP5D with CRC yet, which in the other hand, necessitates further exploration of that gene.

Table 4.4: Pairwise interactions between 44 significant SNPs

SNP1 (Gene1)	SNP2 (Gene2)	IG (%)	P-value
rs2010907	rs3760948 (ARRDC5)	1.30	0.002
rs4625115	rs344570 (TNFSF14)	1.07	0.004
rs9688110 (FAT2)	rs658836	1.12	0.005
rs11185516 (ZDHHC19)	rs344570 (TNFSF14)	1.15	0.008
rs10814848 (GLIS3)	rs6578849 (SYT9)	0.98	0.008
rs11185516 (ZDHHC19)	rs3842986	1.01	0.010
rs1505229 (LRRTM4)	rs952880 (KCNQ5)	0.98	0.011
rs11783793	rs11610311	0.93	0.012
rs1505229 (LRRTM4)	rs11610311	0.98	0.013
rs9288684 (INPP5D)	rs11985944	0.70	0.015
rs9288684 (INPP5D)	rs2179321 (PLCB4)	1.04	0.015
rs1367128 (THSD7B)	rs8022574	1.02	0.016
rs12695485 (LOC107986044)	rs898438 (DCC)	0.79	0.016
rs721619 (EPHX2)	rs4961513	0.92	0.017
rs4625115	rs2571219 (ATP8B1)	0.91	0.017
rs9288684 (INPP5D)	rs2386946 (CDH4)	0.96	0.018
rs1991915 (OTOP1)	rs3842986	0.95	0.019

Moreover, we also calculate three-way IG between the 44 SNPs for $\binom{44}{3}$ times. This time, we set the (optional) threshold of $p < 0.001$, i.e., 1 time out of 1000 permutations, to only keep interactions with significant p-value, that result in 16

significant three-way interactions. Table 4.5 shows the SNPs, IG value, and p-value (which is the significance of IG comparing to 1000 times of permutation) for these 16 pairs. Next to each SNP its corresponding gene is shown to indicate the interactions among the genes. In Table 4.5 we see that genes ALK, DCC, FAT2, and NID2 appear to have significant three-way interactions in association with CRC. In particular, the three-way interaction of ALK, JPH1, and DCC, with IG of 1.93% and p-value of 0.001, more than ever concedes the significance of genes ALK and DCC to be associated with CRC.

Table 4.5: Three-way interactions between 44 significant SNPs

SNP1 (Gene1)	SNP2 (Gene2)	SNP3 (Gene3)	IG (%)	P-value
rs647831	rs2736486	rs952880 (KCNQ5)	2.23	0
rs1816647	rs10814848 (GLIS3)	rs3760948 (ARRDC5)	2.43	0
rs7594717 (ALK)	rs721619 (EPHX2)	rs3760948 (ARRDC5)	2.25	0
rs12695485 (LOC107)	rs17831158 (LINC)	rs13263313 (JPH1)	2.25	0
rs11185516 (ZDHHC)	rs2736486	rs10814848 (GLIS3)	2.11	0
rs11185516 (ZDHHC)	rs11985944	rs3844138	2.41	0
rs1991915 (OTOP1)	rs721619 (EPHX2)	rs17831158 (LINC)	2.55	0
rs1991915 (OTOP1)	rs8022574	rs2571219 (ATP8B1)	2.22	0
rs2010907	rs9688110 (FAT2)	rs4625115	2.27	0
rs12407198 (C1orf101)	rs10016091 (SCFD2)	rs2010907	1.86	0.001
rs1816647	rs6782709 (LOC105)	rs4961513	1.82	0.001
rs7594717 (ALK)	rs13263313 (JPH1)	rs898438 (DCC)	1.93	0.001
rs1367128 (THSD7B)	rs17831158 (LINC)	rs2386946 (CDH4)	1.89	0.001
rs1367128 (THSD7B)	rs6578849 (SYT9)	rs344570 (TNFSF14)	1.40	0.001
rs17379465	rs2645737 (NID2)	rs658836	1.90	0.001
rs4961513	rs11610311	rs3842986	1.82	0.001

4.4.4 Discussion

As discussed before, revealing causality of inheritable diseases through detecting interactions between genetic markers and phenotypes is a difficult task. In addition, the genetic variants discovered in GWAS account for only a small fraction of the

phenotypic variations due to the fact that most effects are expected to be small [49]. Given this fact, the exploration of the inherited genetic variants in common diseases using machine learning methods is the current best approach.

In this regard, we utilized two ensemble learning algorithms, which are known to be able to detect interactions between variables, in order to detect significant interacting genetic markers. Different parameter settings of RF and GBM methods were applied to the CRC dataset in order to find the setting, which produce highest classification accuracy. The best parameter setting was then used to detect the most important SNPs affecting the disease status. From the comparison of the results of the ensemble methods, 44 most significant SNPs, which considered important by both algorithms, were selected for further analysis. The biological interpretations of these SNPs using online databases found 29 corresponding genes. Amongst these genes, DCC and ALK have been shown to have association with CRC based on numerous studies. Consecutively, the enrichment analysis of the genes revealed biological pathways such as 'tobacco use disorder' and 'cell migration'.

Moreover, pairwise and three-way interaction analysis between 44 important SNPs revealed significant interactions between those SNPs, specifically the interaction of DCC and LOC107986044, which can be quite significant. Gene DCC appeared three times as significantly having association with CRC: from the results of ensemble learning method as having significant main effect on disease status, and from the interaction analysis in the pairwise and three-way interactions having significant IG and p-value.

The results of this study showed that ensemble algorithms are powerful approaches for analyzing GWAS data. As illustration, the gene DCC which was shown to be

greatly associated with CRC. However, the ensemble methods have drawbacks in the sense that they cannot handle GWAS high dimensional data containing huge number of genetic markers. Another shortcoming posed to this study is the small size of samples in the dataset, which can produce unreliable results when conducting a GWA study. We resolved these issues by reducing the size of feature set using feature selection methods.

Chapter 5

Discussion

5.1 Summary

The goal of GWAS is to identify genetic markers that can explain complex human diseases. Most existing analyses for GWAS look at one gene at time due to the limitation of analytical methodologies and computational resources. Such a strategy very likely overlook potentially important genetic attributes that have low main effects but contribute to a disease outcome through multifactorial interactions. Detecting such non-additive gene-gene interactions help us better understand the underlying genetic background of common diseases and effectively develop new strategies to treat, diagnose, and prevent them.

Detecting gene-gene interactions for GWAS imposes computational challenges since enumerating combinations of genetic attributes becomes inhibitive when up to a million variables are under consideration. Thus, feature selection becomes a necessity for the task. In addition, utilizing appropriate computational methods capable

of detecting those interactions is another need for a GWA study.

In this thesis, we did a whole genome study in which a GWAS dataset was undergone quality control steps, reduced in size by feature selection methods, and investigated by computational methods to detect significant variations associating with the CRC disease. The results were then validated through biological databases. We performed four primary steps in order to accomplish a GWA study on the CRC data.

Numerous quality control steps were applied to the raw CRC genetic dataset to remove sub-standard samples and low-quality genetic variants. All of the QC steps were conducted using PLINK command-line tool. The original CRC dataset before QC had more than 250 thousands SNPs for 1152 samples. After QC and removing inconsistency in the dataset, the size of the dataset was reduced to 186,151 SNPs for 944 samples.

Even though QC refined the dataset to some extent, this size of feature set still imposed burden for the computational methods. However, some ML methods may be able to handle this size, but the results of these methods would not be reliable and robust because of the curse of dimensionality. Therefore, we did a thorough comparison of six feature selection methods to determine the method which can better detect significant SNPs in the dataset in term of interactions among SNPs. Three univariate feature selection methods (logistic regression, chi-square, odds-ratio) and three multi-variate feature selection methods (ReliefF, TuRF, SURF) were applied to a simulated dataset and the CRC dataset. In the simulated dataset, the performance of FS methods were compared based on their ability to identify and rank the existing interacting SNPs. In the CRC dataset, the methods which detected SNPs demonstrating higher values of interactions (information gain) considered most significant

than the others. The comparison of these methods showed that TuRF outperformed other methods both in simulated and real dataset.

Subsequently, TuRF feature selection method was applied to reduce the size of the dataset to 2,798 most significant SNPs. Two ensemble algorithms, Random Forests and Gradient Boosting Machine, were applied to detect significant main effects and interactions among SNPs. Different configurations of the hyper-parameters of these two methods were applied to identify the parameters setting which produce the highest performance evaluation. The best values for hyper-parameters of the ensemble methods were used to identify the significant SNPs contributing to the disease status. Only the common SNPs from the results of these two methods were extracted for further investigation because of their significance by both methods.

From the results of computational analysis, 44 SNPs were detected as the most significant genetic markers in the CRC dataset. These SNPs were explored in the biological databases for identifying the corresponding genes associated with them. Out of 44 SNPs, 29 genes were found to be associated with these SNPs in which of greatest importance are genes DCC, ALK, ITGA1, E2F3, and NID2 that are known to be associated with CRC based on numerous studies. Enrichment analysis of these 29 genes showed biological pathways such as *tobacco disorder disease* associating with the CRC disease. Furthermore, the pairwise and three-order interaction analysis of the 44 SNPs revealed significant interactions in association with CRC such as the interaction between ALK, JPH1, and DCC.

5.2 Impact

The results of this study showed that the ensemble algorithms are a powerful tool for detecting interactions between SNPs in a genetic dataset. In addition, the detected genetic markers from the results of these methods can be investigated and used in order to prevent or treat the disease. From the analysis of the 44 most significant detected SNPs, 29 associating genes were found to be related to the CRC. Amongst them, five genes are already known to be associated with CRC while others still need further investigations.

5.3 Future Work

In future studies, we expect to explore more sophisticated feature selection algorithms, especially wrapper and embedded methods, and test their utilities in genetic association and bioinformatics studies. In addition, a comparative study can be conducted by including more ML methods in order to obtain robust and reasonable conclusions. Furthermore, network-based analysis which is a novel approach in GWAS can be used for the investigation of the CRC genetic data. For the identified genes, or for the regions of chromosomes 2 and 8 that found to be important (based on Figure 4.7), a candidate gene (association) study, which serve to validate findings from GWAS as well as further explore the biological and clinical interactions between genes, can be conducted to gain deep understanding of their association with the disease phenotype. Moreover, we can develop an application with graphical user interface (GUI) for other researchers to adopt the methods used in this study.

5.4 Conclusion

We conducted a thorough GWA study in which all of the required steps were investigated and performed on a real dataset. We used feature selection methods to reduce the size of the CRC dataset and utilized ensemble algorithms to detect significant interactions between SNPs. The ensemble methods successfully detected strong interacting SNPs which resulted in identifying 44 significant SNPs. The biological interpretation of these 44 SNPs found 29 genes to be associated with the CRC. Moreover, the enrichment analysis of these genes revealed a biological pathway associated with the CRC phenotype. Contributions of this study are manifold such that important genetic variants, associating genes, and biological pathways relating to CRC were detected. Moreover, the capability of ensemble algorithms in the context of GWAS in analyzing bioinformatics data for association studies was elucidated.

Bibliography

[1] D. L. Aisner, T. T. Nguyen, D. D. Paskulin, A. T. Le, J. Haney, N. Schulte, F. Chionh, J. Hardingham, J. Mariadason, N. Tebbutt, et al. Ros1 and alk fusions in colorectal cancer, with evidence of intratumoral heterogeneity for molecular drivers. *Molecular Cancer Research*, 12(1):111–118, 2014.

[2] Y. Akao, S. Noguchi, A. Iio, K. Kojima, T. Takagi, and T. Naoe. Dysregulation of microrna-34a expression causes drug-resistance to 5-fu in human colon cancer dld-1 cells. *Cancer letters*, 300(2):197–204, 2011.

[3] C. A. Anderson, F. H. Pettersson, G. M. Clarke, L. R. Cardon, A. P. Morris, and K. T. Zondervan. Data quality control in genetic case-control association studies. *Nature protocols*, 5(9):1564–1573, 2010.

[4] D. J. Balding. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10):781–791, 2006.

[5] A. L. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97(1):245–271, 1997.

[6] S. Boudjadi, J. Carrier, J. Groulx, and J. Beaulieu. Integrin $\alpha 1\beta 1$ expression is controlled by c-myc in colorectal cancer cells. *Oncogene*, 35(13):1671–1678, 2016.

[7] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[8] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[9] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *Journal of machine learning research*, 13(Jan):27–66, 2012.

[10] A. Bureau, J. Dupuis, K. Falls, K. L. Lunetta, B. Hayward, T. P. Keith, and P. Van Eerdewegh. Identifying snps predictive of phenotype using random forests. *Genetic epidemiology*, 28(2):171–182, 2005.

[11] W. S. Bush and J. H. Moore. Genome-wide association studies. *PLoS computational biology*, 8(12):e1002822, 2012.

[12] M. Castets, L. Broutier, Y. Molin, M. Brevet, G. Chazot, N. Gadot, A. Paquet, L. Mazelin, L. Jarrosson-Wuilleme, J.-Y. Scoazec, et al. Dcc constrains tumour progression via its dependence receptor activity. *Nature*, 482(7386):534–537, 2012.

[13] C. Consortium et al. A comprehensive 1000 genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature genetics*, 47(10):1121–1130, 2015.

[14] W. T. C. C. Consortium et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661, 2007.

[15] N. R. Cook, R. Y. Zee, and P. M. Ridker. Tree and spline based association analysis of gene–gene interaction models for ischemic stroke. *Statistics in medicine*, 23(9):1439–1453, 2004.

[16] H. J. Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human molecular genetics*, 11(20):2463–2468, 2002.

[17] T. M. Cover and J. A. Thomas. *Elements of Information Theory: Second Edition*. Wiley, 2006.

[18] G. M. D'Angelo, D. Rao, and C. C. Gu. Combining least absolute shrinkage and selection operator (lasso) and principal-components analysis for detection of gene-gene interactions in genome-wide association studies. In *BMC proceedings*, volume 3, page S62. BioMed Central, 2009.

[19] M. Dash and H. Liu. Feature selection for classification. *Intelligent data analysis*, 1(1-4):131–156, 1997.

[20] G. Dennis, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki. David: database for annotation, visualization, and integrated discovery. *Genome biology*, 4(9):R60, 2003.

[21] T. G. Dietterich et al. Ensemble methods in machine learning. *Multiple classifier systems*, 1857:1–15, 2000.

[22] D. F. Easton and R. A. Eeles. Genome-wide association studies in cancer. *Human Molecular Genetics*, 17(R2):R109–R115, 2008.

[23] R. Fan, M. Zhong, S. Wang, Y. Zhang, A. Andrew, M. Karagas, H. Chen, C. I. Amos, M. Xiong, and J. H. Moore. Entropy-based information gain approaches to detect and to characterize gene-gene and gene-environment interactions/correlations of complex diseases. *Genetic Epidemiology*, 35(7):706–721, 2011.

[24] A. S. Foulkes. *Applied statistical genetics with R: for population-based association studies*. Springer Science & Business Media, 2009.

[25] K. A. Frazer, S. S. Murray, N. J. Schork, and E. J. Topol. Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, 10(4):241–251, 2009.

[26] A. A. Freitas. *Data mining and knowledge discovery with evolutionary algorithms*. Springer Science & Business Media, 2013.

[27] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[28] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[29] M. García-Magariños, I. López-de Ullibarri, R. Cao, and A. Salas. Evaluating the ability of tree-based methods and logistic regression for the detection of snp-snp interaction. *Annals of human genetics*, 73(3):360–369, 2009.

[30] R. A. Gibbs, J. W. Belmont, P. Hardenbol, T. D. Willis, F. Yu, H. Yang, L.-Y. Ch'ang, W. Huang, B. Liu, Y. Shen, et al. The international hapmap project. *Nature*, 426(6968):789–796, 2003.

[31] B. A. Goldstein, A. E. Hubbard, A. Cutler, and L. F. Barcellos. An application of random forests to a genome-wide association dataset: methodological considerations & new findings. *BMC genetics*, 11(1):1, 2010.

[32] C. S. Greene, N. M. Penrod, J. Kiralis, and J. H. Moore. Spatially uniform relieff (surf) for computationally-efficient filtering of gene-gene interactions. *BioData mining*, 2(1):5, 2009.

[33] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[34] T. Hastie, R. Tibshirani, and J. Friedman. The elements of statistical learning: Data mining, inference, and prediction. *Biometrics*, 2002.

[35] A. G. Heidema, J. M. Boer, N. Nagelkerke, E. C. Mariman, E. J. Feskens, et al. The challenge for genetic epidemiologists: how to analyze large numbers of snps in relation to complex diseases. *BMC genetics*, 7(1):23, 2006.

[36] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio. Potential etiologic and functional implications of

genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, 2009.

[37] J. N. Hirschhorn and M. J. Daly. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108, 2005.

[38] T. Hu, Y. Chen, J. W. Kiralis, R. L. Collins, C. Wejse, G. Sirugo, S. M. Williams, and J. H. Moore. An information-gain approach to detecting three-way epistatic interactions in genetic association studies. *Journal of the American Medical Informatics Association*, 20(4):630–636, 2013.

[39] J. Hua, W. D. Tembe, and E. R. Dougherty. Performance of feature-selection methods in the classification of high-dimension data. *Pattern Recognition*, 42(3):409–424, 2009.

[40] D. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37(1):1–13, 2008.

[41] P. Jia and Z. Zhao. Network-assisted analysis to prioritize gwas results: principles, methods and perspectives. *Human genetics*, 133(2):125–138, 2014.

[42] Y. Kim, R. Wojciechowski, H. Sung, R. A. Mathias, L. Wang, A. P. Klein, R. K. Lenroot, J. Malley, and J. E. Bailey-Wilson. Evaluation of random forests performance for genome-wide association studies in the presence of interaction effects. In *BMC proceedings*, volume 3, page S64. BioMed Central, 2009.

[43] K. Kira and L. A. Rendell. A practical approach to feature selection. In *Proceedings of the ninth international workshop on Machine learning*, pages 249–256, 1992.

[44] R. J. Klein, C. Zeiss, E. Y. Chew, J.-Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, et al. Complement factor h polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389, 2005.

[45] I. Kononenko. Estimating attributes: analysis and extensions of relief. In *European conference on machine learning*, pages 171–182. Springer, 1994.

[46] H. Kunugi, R. Hashimoto, T. Okada, H. Hori, T. Nakabayashi, A. Baba, K. Kudo, M. Omori, S. Takahashi, R. Tsukue, et al. Possible association between nonsynonymous polymorphisms of the anaplastic lymphoma kinase (alk) gene and schizophrenia in a japanese population. *Journal of neural transmission*, 113(10):1569–1573, 2006.

[47] G. Lettre and J. D. Rioux. Autoimmune diseases: insights from genome-wide association studies. *Human molecular genetics*, 17(R2):R116–R121, 2008.

[48] H. Li, Y. Lee, J. L. Chen, E. Rebman, J. Li, and Y. A. Lussier. Complex-disease networks of trait-associated single-nucleotide polymorphisms (SNPs) unveiled by information theory. *Journal of American Medical Informatics Association*, 19:295–305, 2012.

[49] P. Lichtenstein, N. V. Holm, P. K. Verkasalo, A. Iliadou, J. Kaprio, M. Koskenvuo, E. Pukkala, A. Skytthe, and K. Hemminki. Environmental and heritable

factors in the causation of cancer-analyses of cohorts of twins from sweden, denmark, and finland. *New England journal of medicine*, 343(2):78–85, 2000.

[50] D. Lipson, M. Capelletti, R. Yelensky, G. Otto, A. Parker, M. Jarosz, J. A. Curran, S. Balasubramanian, T. Bloom, K. W. Brennan, et al. Identification of new alk and ret gene fusions from colorectal and lung cancer biopsies. *Nature medicine*, 18(3):382–384, 2012.

[51] G. Lubke, C. Laurin, R. Walters, N. Eriksson, P. Hysi, T. Spector, G. Montgomery, N. Martin, S. Medland, and D. Boomsma. Gradient boosting as a snp filter: An evaluation using simulated and hair morphology data. *Journal of data mining in genomics & proteomics*, 4, 2013.

[52] S. Ma and J. Huang. Penalized feature selection and classification in bioinformatics. *Briefings in bioinformatics*, 9(5):392–403, 2008.

[53] 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*, 45(D1):D896–D901, 2017.

[54] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. Ioannidis, and J. N. Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews genetics*, 9(5):356–369, 2008.

[55] K. Michailidou, P. Hall, A. Gonzalez-Neira, M. Ghoussaini, J. Dennis, R. L. Milne, M. K. Schmidt, J. Chang-Claude, S. E. Bojesen, M. K. Bolla, et al.

Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nature genetics*, 45(4):353–361, 2013.

[56] K. L. Mohlke, M. Boehnke, and G. R. Abecasis. Metabolic and cardiovascular traits: an abundance of recently identified common genetic variants. *Human molecular genetics*, 17(R2):R102–R108, 2008.

[57] J. H. Moore. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human heredity*, 56(1-3):73–82, 2003.

[58] J. H. Moore, F. W. Asselbergs, and S. M. Williams. Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, 26(4):445–455, 2010.

[59] J. H. Moore and M. D. Ritchie. The challenges of whole-genome approaches to common diseases. *Jama*, 291(13):1642–1643, 2004.

[60] J. H. Moore and B. C. White. Tuning relieff for genome-wide genetic analysis. In *European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, pages 166–175. Springer, 2007.

[61] J. H. Moore and S. M. Williams. New strategies for identifying gene-gene interactions in hypertension. *Annals of medicine*, 34(2):88–95, 2002.

[62] J. H. Moore and S. M. Williams. Epistasis and its implications for personal genetics. *The American Journal of Human Genetics*, 85(3):309–320, 2009.

[63] R. S. Olson, W. La Cava, Z. Mustahsan, A. Varik, and J. H. Moore. Data-driven advice for applying machine learning to bioinformatics problems. *arXiv preprint arXiv:1708.05070*, 2017.

[64] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[65] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[66] D. E. Reich and E. S. Lander. On the allelic spectrum of human disease. *TRENDS in Genetics*, 17(9):502–510, 2001.

[67] D. M. Reif, A. A. Motsinger, B. A. McKinney, J. E. Crowe, and J. H. Moore. Feature selection using a random forests classifier for the integrated analysis of multiple data types. In *Computational Intelligence and Bioinformatics and Computational Biology, 2006. CIBCB'06. 2006 IEEE Symposium on*, pages 1–8. IEEE, 2006.

[68] G. Ridgeway et al. gbm: Generalized boosted regression models. *R package version*, 1(3):55, 2006.

[69] M. D. Ritchie, L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl, and J. H. Moore. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *The American Journal of Human Genetics*, 69(1):138–147, 2001.

[70] M. Robnik-Šikonja and I. Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine learning*, 53(1-2):23–69, 2003.

[71] Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517, 2007.

[72] N. J. Samani, J. Erdmann, A. S. Hall, C. Hengstenberg, M. Mangino, B. Mayer, R. J. Dixon, T. Meitinger, P. Braund, H.-E. Wichmann, et al. Genomewide association analysis of coronary artery disease. *New England Journal of Medicine*, 357(5):443–453, 2007.

[73] F. R. Schumacher, S. L. Schmit, S. Jiao, C. K. Edlund, H. Wang, B. Zhang, L. Hsu, S.-C. Huang, C. P. Fischer, J. F. Harju, et al. Genome-wide association study of colorectal cancer identifies six new susceptibility loci. *Nature communications*, 6:7138, 2015.

[74] D. F. Schwarz, S. Szymczak, A. Ziegler, and I. R. König. Picking single-nucleotide polymorphisms in forests. In *BMC proceedings*, volume 1, page S59. BioMed Central, 2007.

[75] S. C. Shah and A. Kusiak. Data mining and genetic algorithm based gene/snp selection. *Artificial intelligence in medicine*, 31(3):183–196, 2004.

[76] B. Stewart, C. P. Wild, et al. World cancer report 2014. *Health*, 2017.

[77] Y. V. Sun, L. F. Bielak, P. A. Peyser, S. T. Turner, P. F. Sheedy, E. Boerwinkle, and S. L. Kardia. Application of machine learning algorithms to predict

coronary artery calcification with a sibship-based design. *Genetic epidemiology*, 32(4):350–360, 2008.

[78] Y. V. Sun, Z. Cai, K. Desai, R. Lawrance, R. Leff, A. Jawaid, S. L. Kardia, and H. Yang. Classification of rheumatoid arthritis status with candidate gene and genome-wide single-nucleotide polymorphisms using random forests. In *BMC proceedings*, volume 1, page S62. BioMed Central, 2007.

[79] Y. V. Sun, K. A. Shedden, J. Zhu, N.-H. Choi, and S. L. Kardia. Identification of correlated genetic variants jointly associated with rheumatoid arthritis using ridge regression. In *BMC proceedings*, volume 3, page S67. BioMed Central, 2009.

[80] M. Szumilas. Explaining odds ratios. *Journal of the Canadian academy of child and adolescent psychiatry*, 19(3):227, 2010.

[81] 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*, 33(S1):S51–S57, 2009.

[82] R. Tang, J. P. Sinnwell, J. Li, D. N. Rider, M. de Andrade, and J. M. Biernacka. Identification of genes and haplotypes that predict rheumatoid arthritis using random forests. In *BMC proceedings*, volume 3, page S68. BioMed Central, 2009.

[83] R. Upstill-Goddard, D. Eccles, J. Fliege, and A. Collins. Machine learning approaches for the discovery of gene–gene interactions in disease data. *Briefings in bioinformatics*, 14(2):251–260, 2012.

[84] R. J. Urbanowicz, J. Kiralis, N. A. Sinnott-Armstrong, T. Heberling, J. M. Fisher, and J. H. Moore. Gametes: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures. *BioData mining*, 5(1):16, 2012.

[85] R. J. Urbanowicz, J. W. Kiralis, J. M. Fisher, and J. H. Moore. Predicting the difficulty of pure, strict, epistatic models: metrics for simulated model selection. *BioData Mining*, 5:15, 2012.

[86] S. Van Slambrouck, C. Grijelmo, O. De Wever, E. Bruyneel, S. Emami, C. Gespach, and W. F. Steelant. Activation of the fak-src molecular scaffolds and p130cas-jnk signaling cascades by α1-integrins during colon cancer cell invasion. *International journal of oncology*, 31(6):1501–1508, 2007.

[87] 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*, 291(5507):1304–1351, 2001.

[88] M. Wang, X. Chen, M. Zhang, W. Zhu, K. Cho, and H. Zhang. Detecting significant single-nucleotide polymorphisms in a rheumatoid arthritis study using random forests. In *BMC proceedings*, volume 3, page S69. BioMed Central, 2009.

[89] W. Y. Wang, B. J. Barratt, D. G. Clayton, and J. A. Todd. Genome-wide association studies: theoretical and practical concerns. *Nature Reviews Genetics*, 6(2):109–118, 2005.

[90] P. Watson, R. Ashwathnarayan, H. T. Lynch, and H. K. Roy. Tobacco use and increased colorectal cancer risk in patients with hereditary nonpolyposis colorectal cancer (lynch syndrome). *Archives of internal medicine*, 164(22):2429–2431, 2004.

[91] M. N. Wright and A. Ziegler. ranger: A fast implementation of random forests for high dimensional data in c++ and r. *arXiv preprint arXiv:1508.04409*, 2015.

[92] M. N. Wright, A. Ziegler, and I. R. König. Do little interactions get lost in dark random forests? *BMC bioinformatics*, 17(1):1, 2016.

[93] B. Xue, M. Zhang, W. N. Browne, and X. Yao. A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on Evolutionary Computation*, 20(4):606–626, 2016.

[94] C. Yang, Z. He, X. Wan, Q. Yang, H. Xue, and W. Yu. Snpharvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies. *Bioinformatics*, 25(4):504–511, 2009.

[95] W. W. Yang and C. C. Gu. Selection of important variables by statistical learning in genome-wide association analysis. In *BMC proceedings*, volume 3, page S70. BioMed Central, 2009.

[96] F. Yates. Contingency tables involving small numbers and the χ 2 test. *Supplement to the Journal of the Royal Statistical Society*, 1(2):217–235, 1934.

[97] L. Yu and H. Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *ICML*, volume 3, pages 856–863, 2003.

[98] B. W. Zanke, C. M. Greenwood, J. Rangrej, R. Kustra, A. Tenesa, S. M. Farrington, J. Prendergast, S. Olschwang, T. Chiang, E. Crowdy, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nature genetics*, 39(8):989–994, 2007.

[99] R. Zhang, M. Chu, Y. Zhao, C. Wu, H. Guo, Y. Shi, J. Dai, Y. Wei, G. Jin, H. Ma, et al. A genome-wide gene–environment interaction analysis for tobacco smoke and lung cancer susceptibility. *Carcinogenesis*, 35(7):1528–1535, 2014.

[100] A. Ziegler, A. L. DeStefano, and I. R. König. Data mining, neural nets, trees–problems 2 and 3 of genetic analysis workshop 15. *Genetic epidemiology*, 31(S1), 2007.