Genome-wide association study of colorectal cancer using evolutionary computing

by

© *Shengkai Geng*

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

Jan 2021

St. John's Newfoundland

Abstract

The heritability of complex diseases is usually ascribed to interacting genetic alterations. Many diseases have been found that are influenced by genetic factors. Colorectal cancer (CRC) is a type of cancer starting from the colon or rectum that seriously threatens human health, and it has the chance to spread to other parts of the human body. The cause of CRC is multifactorial, including age, sex, intake of fat, etc. In addition, it has been suggested that genetic factors also play an essential role. Several genetic variations have been identified as associated with CRC. However, they only explain a small portion of the heritability. More advanced computational techniques are required to identify combinations of genetic factors. Recently, artificial intelligence algorithms have became a powerful tool for biomedical data analyses. In this thesis, I design an evolutionary algorithm for the identification of combinations of genetic factors, i.e., single nucleotide polymorphisms (SNPs), that can best explain the susceptibility to CRC.

Acknowledgements

First of all, I give my thanks to my supervisor, Dr. Ting Hu. Without her help, I could not have finished this thesis. Even when I made mistakes time and time again, she was always patient, supporting me, and giving me much inspiration to solve problems. I really appreciate that.

I want to thanks all my friends. They have been extremely helpful to me throughout time. I gained a lot from them, and I will never forget.

I want to thank my parents, especially my mother, who has given me tremendous support during this period. There no word that can express my thanks. I am grateful.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Many human diseases have been proven to pass from generation to generation [2, 3]. Abnormal genes in a specific position can lead to a high risk to contract a disease during a lifetime or directly cause a disease phenotype [4, 5, 6]. Recent study has proved that some diseases can be caused by a single mutation at a single gene [7]. Therefore, having a good understanding of genes can help make better predictions, diagnose, treat, and prevent a variety of diseases [8, 9, 10].

Understanding complex diseases' genetic etiology is challenging [11]. Many diseases have a complex genetic architecture. A disease phenotype could be influenced by a large number of genes collectively [12]. Finding genetic variants that affect a disease not only requires investigation of independent genes but also detection of their interactions. Hence, more powerful methods to study the genes associated with diseases are needed.

Genome-wide association studies (GWAS) are observational studies that investigate genetic variants across the genomes by analyzing population-based data of cases

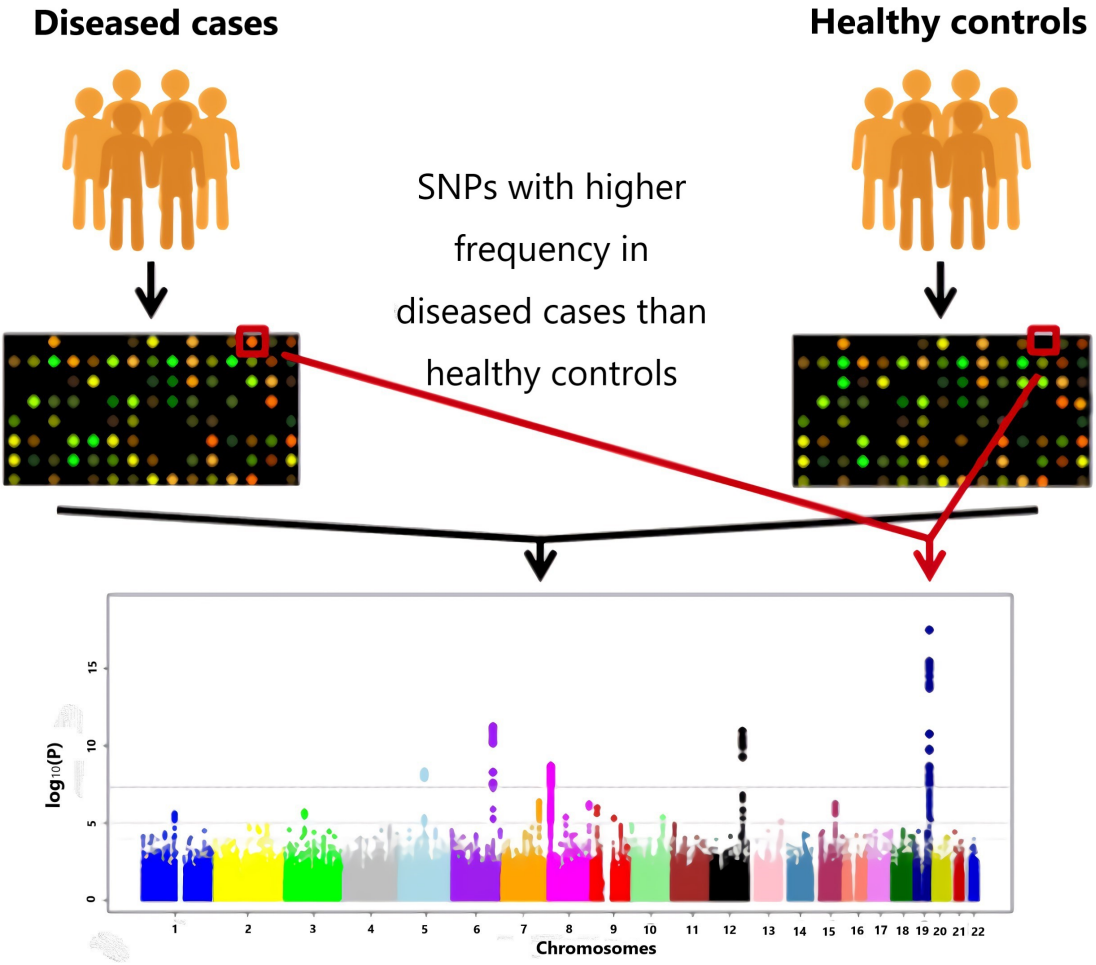


Figure 1.1: Genome-wide association studies.

GWAS analyze common variants in cases and controls to find variants associated with a disease. Then statistical analysis is usually used to test how likely a genetic variant is to be associated with a disease [13].

and controls to identify genotype-phenotype associations [14]. The primary purpose of GWAS is to identify the associations between single-nucleotide polymorphisms (SNPs) and traits like major human diseases by scanning the entire set of genetic variants in different individuals. Fig.1.1 shows the procedures of GWAS. GWAS have revolutionized the field of complex disease genetics over the past decade, providing numerous compelling associations for complex human traits and diseases [1].

GWAS aim to find SNPs with the strongest association with a disease. Research have proven that colorectal cancer is affected by multiple SNPs instead of one single SNP [15]. The interactions between SNPs are important to analyze and have significant genetic meaning. Therefore, finding disease-associated genes can be considered as a data selection problem. Data selection aims to find the optimum samples or features for the learning method to achieve the best performance [16, 17].

Machine learning, as a branch of artificial intelligence, provides efficient ways to analyze a large amount of data. Machine learning algorithms build a mathematical model of sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task [18]. With the rapid development of numerical methods in recent years, increasingly efficient approaches have helped researchers to solve complex problems in many areas. Machine learning methods for analyzing big data can help reduce costs and save time in the mining process [19]. An increasing number of machine learning methods that help generate multiple associated results in the human system have been devised [20, 21].

Evolutionary computing is a sub-field of artificial intelligence; it is a technique based on the Darwinian principles of natural selection and evolution [22, 23]. The fundamental metaphor of evolutionary computing relates this powerful natural evo-

lution to a particular style of problem-solving [24]. Evolutionary computing can be used in a wide range of problems, and it can produce highly optimized solutions. A great number of applications of evolutionary computing have been developed. Evolutionary computing is widely used in genetics study today and it has been proven that evolutionary algorithms can exhibit a good performance in genome-wide association studies [25, 26]. Evolutionary computing has many advantages compared to other computational methods. Besides its conceptual simplicity, the broad applicability is very impressive [27]. Moreover, the great potential of using knowledge and hybridization with other methods gives evolutionary computing more flexibility.

Genetic algorithm (GA) has become an extraordinary method to solve optimization problems. GA is a large class of evolutionary algorithms that are commonly used to generate high-quality solutions for optimization and search problems [28]. GA can evolve potential solutions of a problem in the population and eventually find an answer capable of solving the problem. Many researchers have applied GA to solve various complex problems [29, 30].

The inheritable disease that we study is colorectal cancer (CRC). It is the development of cancer from the colon or rectum [31]. It is the fourth most worldwide common cause of cancer death, after lung, stomach, and liver cancer [32]. It is the second leading cause of cancer death in women and the third for men [33]. However, our understanding of this disease is still limited. The lack of understanding of diseases significantly limits patients' treatment. Some studies have proved that there are several subtypes of colorectal cancer [34]. The existence of subtypes increases the complexity of studying colorectal cancer. Generally, colorectal cancer can be divided into categories of microsatellite instability (MSI), and microsatellite stability (MSS),

based on its genetic instability [35]. Each subtype displays different pathological and genetic signatures. This makes the treatment of colorectal cancer more difficult and less efficient.

In this thesis, we propose a GA method to analyze a GWAS dataset for colorectal cancer. This thesis's primary contribution is to identify potential SNPs associated with colorectal cancer aross the human genome. Our method is not only to consider feature selection, but also to concurrently involve sample selection. Our research starts by processing the dataset of colorectal cancer to ensure the quality of the dataset. Then we construct the GA method with the optimised parameters. From the results of the GA method, our GA method shows an excellent performance. We also find some close relationships between cases in the dataset, which is helpful to gain insight into the subtypes of colorectal cancer.

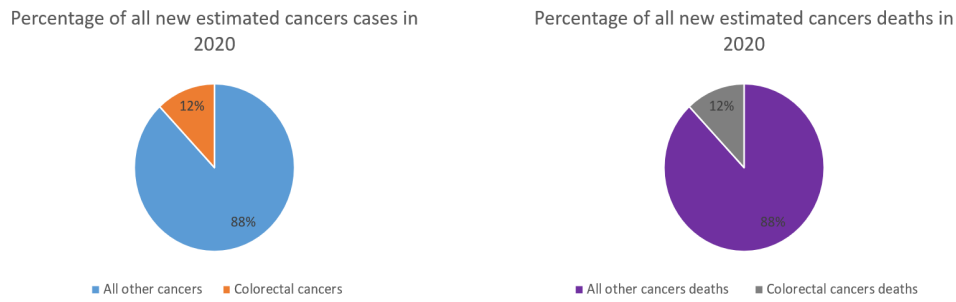
Chapter 2

Background and related work

2.1 Colorectal cancer

In the modern world, more than 1 million people contract colorectal cancer every year. The disease-specific mortality rate is nearly 33% in the developed world. [36]. In 2010, colorectal cancer caused 715,000 deaths, and this number significantly increased compared to 490,000 deaths in 1990 [37]. It can be predicted that this number will continue to keep growing in the future.

Colorectal cancer (CRC) is the third most commonly diagnosed cancer among humans, and it is the second-highest cause of cancer occurrence and death for men and women [38]. Based on the statistics from 2007 to 2009, the rate that people in the US diagnosed with colorectal cancer in their life is 4.96% [39]. In 2017, there were 135,430 individuals projected to be newly diagnosed with CRC and 50,260 deaths from the disease. The mortality of colorectal cancer is still very high in Europe, and it is keep growing in some countries [40]. Fig.2.1 shows that 12% of all estimated



(a) Cases (b) Deaths

Figure 2.1: Colorectal cancer statistics.

The statistics shows that colorectal cancer is one of the most serious cancer nowadays.

new cancer cases are colorectal cancer cases in 2020. Colorectal cancer also accounts for 12% of all cancer deaths. In Japan, the incidence and mortality of colorectal cancer have experienced substantial growth recently. The number of deaths caused by colorectal cancer per unit of population has increased around tenfold during the last 50 years. Mortality due to colorectal cancer is on the rise, surpassing 49,000 in 2015 in Japan [41]. From those statistics, it is easy to understand that colorectal cancer has become a serious problem in the world.

From the recent study of colorectal cancer epidemiology, there are many factors that can be seen as risk factors for colorectal cancer, including older age, male sex, high intake of fat, sugar, alcohol, and lack of exercise [42]. The risk of colorectal cancer diagnosis increases after the age of 40, and rises sharply after age 50 [43]. High fat intake, especially animal fat, is a major dietary risk factor of for colorectal cancer [44]. People who lack physical activity have a higher risk of colorectal cancer than people who are regularly physically active [45]. High consumption of cigarettes or

alcohol also plays an important role in the epidemiology of colorectal cancer [46, 47].

Moreover, a plethora of research has shown that genetic mutation is one of the main factors which can cause CRC [36, 48, 49, 50, 51]. Heredity is one factor of this disease [52], which means it could be passed on from generation to generation. Up to 20% of patients who develop colorectal cancer have other family relatives who have been affected by this disease [53]. Around 5 to 10% of colorectal cancer cases are a consequence of acknowledged hereditary conditions [54]. Five to six percent of the world wide population have a lifetime risk of colorectal cancer in general because of heredity [55]. Hence, to help people, gaining more information about the heredity of this disease is important.

It has been clearly proven that colorectal cancer evolves through multiple pathways to many subtypes [34, 56]. Based on its genetic instability, this disease can be defined as having microsatellite instability (MSI) or microsatellite stability (MSS) [35]. MSI is the condition of genetic hypermutability that results from impaired DNA mismatch repair. MSS cancers are characterized by changes in chromosomal copy number and show worse prognosis. The existence of subtypes increases the complexity of colorectal cancer, and makes treatment more difficult and less efficient.

2.2 Genome-wide association study

With the completion of the human genome project [57] that helped to map the nucleotides contained in a human genome, relevant research gained a great boost in recent decades [58, 59]. The goal of the human genome project is finding the genome pairs that construct human DNA, and identifying and mapping all of the genes of

the human genome from both a physical and a functional standpoint. Genetic markers spanning the whole human genome have empowered widespread mapping efforts based on linkage analysis, using families with several affected individuals, resulting in the discoveries of multiple genes for diseases.

Genome-wide association study (GWAS or GWA study) is one of the most popular observational studies of the genome-wide set of genetic variants in different individuals using single-nucleotide polymorphisms (SNPs), in order to see if there is any association between variants and traits [60, 61, 62]. GWAS is an approach that studies genetic variants by scanning the genome samples of diseased cases and healthy controls [63]. It uses linkage analysis to map genome loci that have an effect on disease or other traits. The primary goal of GWAS is to understand more about biology which will advance better treatment or prevention.

In the past decade, many new observations results made by GWAS have given us more understanding of disease [64]. Fig.2.2 shows the development of GWAS in the past ten years, with increasing discoveries. About 10,000 strong associations between genetic variants and complex traits have been found [1]. GWAS has proved that it is a powerful tool to find the relationship between diseases and human genes [65]. A plethora of research on GWAS has shown that there is a relationship between some traits and the human genome [60, 66]. The Wellcome Trust Case Control Consortium has proven that many SNPs have associations with some common diseases [67]. Their analysis of 16,179 cases and controls was used to study seven common diseases, including diabetes, rheumatoid arthritis, and hypertension.

The most common method of GWAS is the case-control setup, which compares two large groups of individuals, one healthy control group and one case group affected

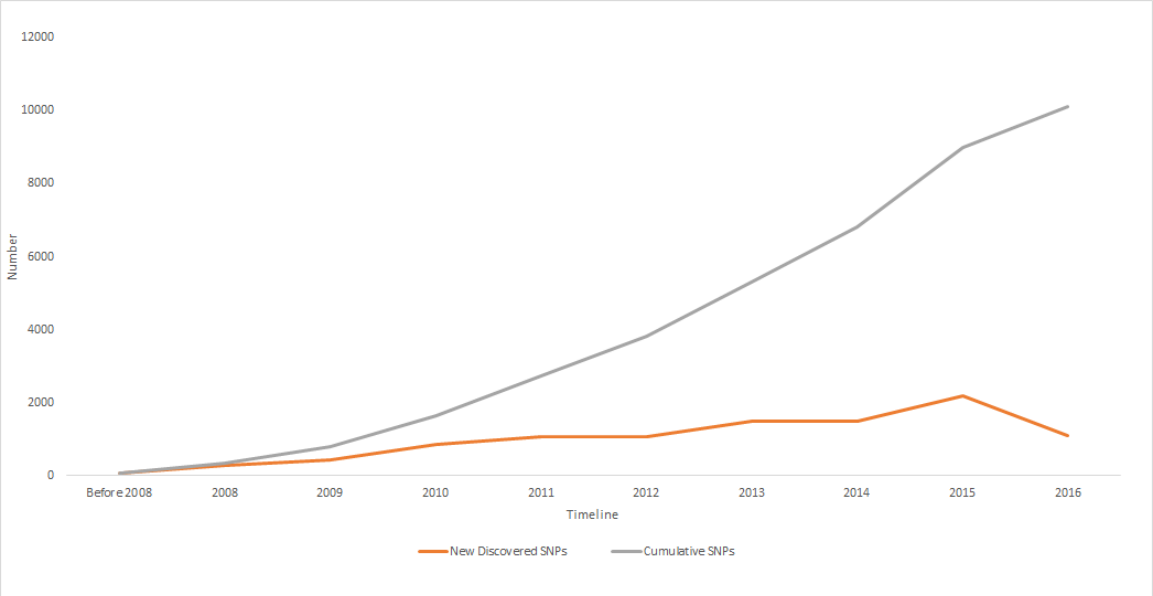


Figure 2.2: GWAS SNP-trait discovery timeline [1].

In the last ten years, the development of GWAS was rapid. Increasing discoveries help to gain more insights from human genes.

by a disease. In each group, every individual is genotyped for the majority of common known SNPs. Then each SNP is investigated to check if allele frequency is significantly altered between the case and the control group.

Although GWAS successfully provides a way to study interactions between genes and diseases, there are still some obstacles. With the impressive development in genomics, the size of genetic data has experienced a rapid increase. With the development of gene study, more dimensions and more samples are being added to biological data. Because of the massive size of the gene data, the study could be very challenging. Overfitting is one of the biggest problems of high dimensional data [68]. The required multiple tests to account for the large number of associations are also a difficulty [69]. Thus, in order to gain more insights from genes, cooperative endeavors are needed with various other research areas, such as computer science and statistics.

2.3 Machine learning

Machine learning, as a branch of artificial intelligence, provides multiple methods to deal with complex problems, such as classification, regression, clusters, etc. Machine learning is the study of computational algorithms that can find the solution to problems by using example data (training data) or experience in order to make predictions or decisions [70]. Based on the existence of labels in the dataset during the learning, machine learning could be divided into three broad categories, supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, the machine is trained with the dataset and each sample's desired outputs to learn the rule that maps inputs to outputs. Unsupervised learning looks for undetected patterns in

a dataset without labels. Reinforcement learning aims to find out what action should be taken in an environment in order to maximize the rewards. The variety of machine learning is very helpful to solve complex problems.

The development of machine learning in the last several decades has been very rapid. Many research areas have applied machine learning methods for their studies. With the development of machine learning in recent decades, the bioinformatic area experienced extensive growth [71, 72]. Moreover, the outcomes of the combination of genome-wide association studies and machine learning have been proven positive [73, 74].

Currently, machine learning methods are widely used for genome-wide association studies. Yang *et al* [75] provided a positive-unlabeled learning algorithm by grouping the dataset in four groups and using a weighted support vector machine to build a classifier to identify general disease genes. The results showed that the performance of their method outperformed three other advanced techniques. They also demonstrated 20 novel disease genes and 8 specific disease classes in total, such as cardiovascular diseases and endocrine diseases. The research of Chun *et al* [76] used a maximum entropy model to filter the dictionary they built to create a system that can automatically extract helpful information, especially the relationship between diseases and genes, from biomedical data sources. It substantially improved precision by 26.7% and slightly decreased the recall of dictionary matching. Maciukiewicz *et al* [77] applied classification-regression trees and a support vector machine method to build predictive models of duloxetine outcomes in a major depressive disorder dataset. They compared the performance of two methods and listed some robust variants across five folds of the nested cross-validation. Han *et al* [78] proposed a method using the

Markov Blanket to detect epistatic interactions in case-control GWAS. This method can be used to detect SNPs that have a strong association with diseases. It also can calculate the association between variables to implement a heuristic search.

Machine learning approaches have provided a great potential for genetics study. However, because the original data usually contains low quality content, using their raw form to deal with large real data is still limited [79]. The conventional machine learning method has also been challenged by the dramatic growth of the dimension size of biological data. The dimensionality of datasets creates considerable difficulty in designing machine learning methods. Algorithms designers have put a great deal of effort into building efficient methods to extract the most suitable content for learning from raw data.

Data selection is one of the main problems in machine learning [80]. It aims to select a subset of data that is highly discriminatory. With data selection, the classification method can achieve a better accuracy [81]. Feature selection helps to simplify models, shorten training time, avoid the curse of dimensionality, and reduce overfitting [82]. Sample selection is also essential for data mining [83]. Compared to feature selection, sample selection receives less attention. The advantages of sample selection are its low cost, closer relationship, and fewer outlying samples. Simultaneous sampling and feature selection has been shown to be an excellent method to deal with classification problems [84].

Using data mining and machine learning methods to explore complex genotype-phenotype relationships is a challenge for GWAS, because associations between genetic variants and traits are usually not very informative [85]. In order to find diseases associated SNPs, GWAS usually needs to deal with a huge number of individuals,

containing many cases and controls. Overfitting and multiple testing are the biggest challenges of GWAS. To deal with a dataset with a great deal of information would be extremely time-consuming and energy-consuming without the help of efficient methods.

2.4 Evolutionary computing

2.4.1 Overview

Evolutionary computing [24, 86] is a subset of machine learning. It is a technique that simulates evolution to find solutions to complex problems [87]. It is a computing method with a special flavour that draws inspiration from the natural evolutionary process. Given a population that contains a number of individuals within some special environment that has limited resources, competition for those resources leads to selection. Those individuals strive for survival and reproduction. Individuals' fitness is the key point that determines if they can survive. The fitter they are, the higher is the chance they can survive and generate the offspring. Computer scientists gained inspiration from this process. Based on this theory, they developed evolutionary computing [88]. In the problem-solving process, as it showed in Fig.2.3, a collection of potential solutions represents the population. Their fitness is the quality of solving the problem. Those solutions with better fitness have a good chance to survive and become parents to pass on their traits. With crossover, individuals with good fitness could generate new offspring that carry their genome, and the new individuals could have higher fitness.

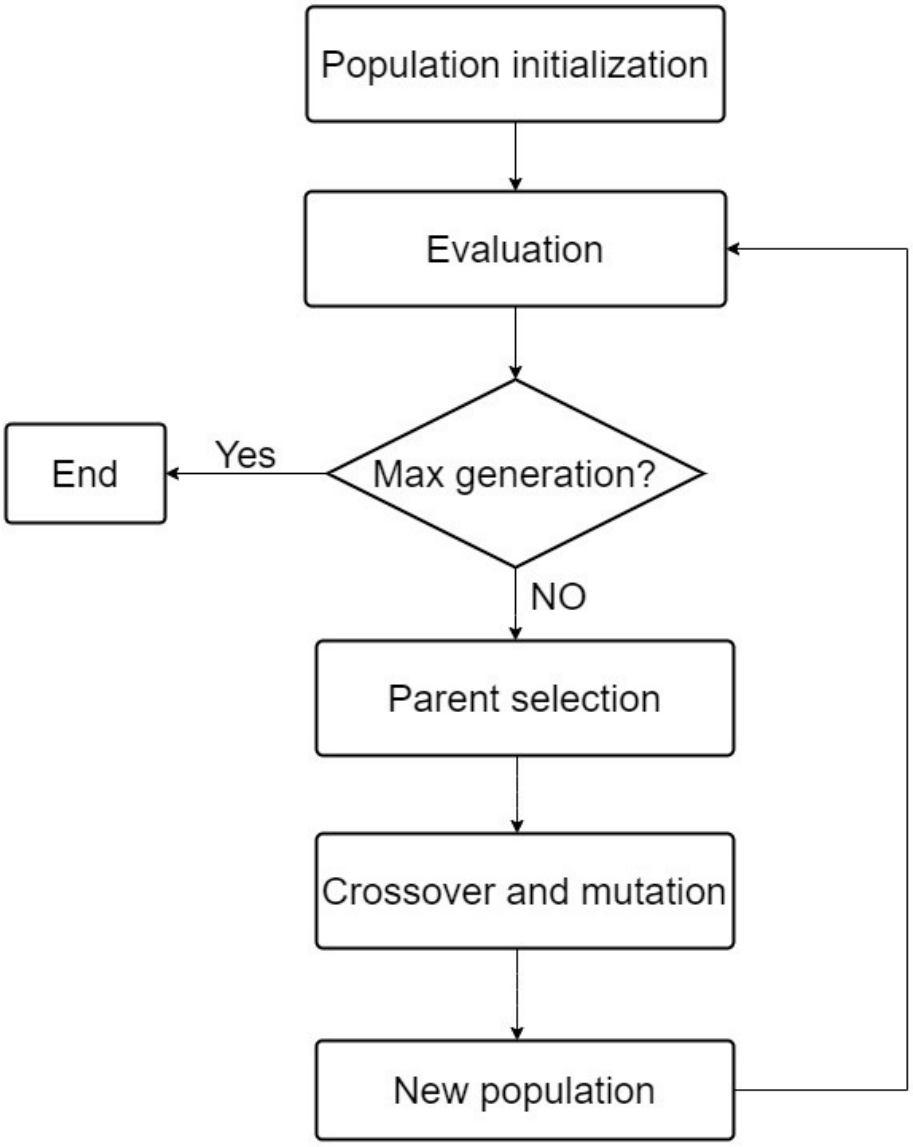


Figure 2.3: Evolutionary computing

Generally, evolution computing contains these steps.

In the last several decades, evolutionary algorithms have been widely used in many areas [89, 90]. In industrial design, Keane *et al* [91] applied the evolutionary algorithm to the design of the case of a satellite dish holder boom that connects the satellite's body with the dish needed for communication. The structure that the evolutionary algorithm generated was almost twenty times better than the traditional structure. In the finance area, Schulenburg *et al* [92] used a learning classifier system based on evolutionary computing to evolve sets of rules to model the behaviour of stock market traders. They used ten years of trading history as the input to evolve trading agents. The results show the evolved trading agents greatly outperformed popular strategies. Compared to other methods such as neural networks, the evolved traders are easier to examine. In the biology area, Eshelman *et al* [93] used an evolutionary algorithm to deal with incest prevention by simulating a known natural phenomenon. The results show that computer-simulated evolution benefits from incest prevention, which strongly confirms that incest brings negative effects to evolutionary processes.

2.4.2 Design of evolutionary algorithm

The effectiveness of an evolutionary algorithm depends on many aspects, such as representation, mutation rate, etc. The parameters are varied, and this plays an important role in evolutionary algorithms. Setting different parameters could lead to different results and performance. Many studies have proven that the performance of an evolutionary algorithm depends very much on the parameters setting [24]. Good parameter setting has a significant influence on the final performance of EA.

The parameters of an evolutionary algorithm can be divided into two types, sym-

bolic parameters and numeric parameters, based on their domains [94]. Symbolic parameters include parameters with a finite domain and no sensible distance metric or ordering, which defines the essence of an evolutionary algorithm, e.g., crossover operator. Numeric parameters are those parameters that are a subset of the real numbers, e.g., population size. This difference has a great influence on searchability. To find the best setting for numerical parameters, heuristic search and optimization methods can be used. For symbolic parameters, there are not many options except sampling.

Evolutionary algorithm design includes all the detailed decisions that are required to be specific. The main challenge of the design process is that every single value of parameters may greatly influence the performance of the evolutionary algorithm. Therefore, the design of an evolutionary algorithm generally is an optimization problem.

Parameter tuning is one of the most common methods to specify values for the evolutionary algorithm parameters, where the parameters are defined before the implementation, and all parameters should remain fixed during the run. Solving the tuning problem is one of the biggest challenges of an evolutionary algorithm. Parameters have the effect of interaction, which means optimizing parameters one by one is not a wise option. Fore a satisfactory performance, specific problems may require an evolutionary algorithm with a specific setup [22].

The most common way to measure the performance of an evolutionary algorithm is through solution quality and algorithm speed. Solution quality can be represented by the fitness function. As for algorithm speed, the search effort needs to be measured, such as the number of fitness evaluations. Hence, the evaluation process can be done

by defining the performance as the best fitness at termination, given a maximum number of generations.

The first step of an evolutionary definition is the representation, which sets up a bridge between the original problem context and the problem-solving space. Representation of an evolutionary algorithm expresses the phenotype of possible solutions to genotypes. To solve an optimization problem, the most popular way is to use an integer string as chromosomes of individuals to be the representation to illustrate candidate solutions.

The fitness function plays a role that expresses the requirements the population should eventually adapt to meet and provides the basis for selection. It is a function that assigns fitness values to all potential solutions and determines how good they are.

The population of an evolutionary algorithm holds potential solutions. Mostly, the size of the population is supposed to be constant and does not change during evolution, in order to create a more competitive environment to generate new individuals with high fitness.

Parent selection allows individuals with better fitness have a higher chance to become parents of the next generation. Individuals with lower fitness are usually assigned a small chance. The role of parent selection is to improve the quality of the population.

Commonly, variation operators include mutation and crossover. Their job is to generate new individuals from old ones, which involves producing new potential solutions. Mutation is a stochastic method that is applied to offspring and slightly changes their genotype. Because mutation is a random and unbiased change, it could

help the learning process to avoid a local optimal solution [95]. Crossover method has to merge two parent individuals' genotypes into one or two offspring. Mostly, the decision of what parts of parents should be combined is random, which also makes crossover a stochastic method.

Because the population size is constant, some individuals should be removed from the population when new offspring are generated. The role of survivor selection is to distinguish among individuals based on their quality and keep the average quality level of the population. Unlike parent selection, which is stochastic, survivor selection usually is deterministic. Thus, survivor selection commonly removes individuals with the lowest fitness.

2.5 Genetic algorithms

Genetic algorithms are the most widely known method that belongs to the class of evolutionary algorithms. Genetic algorithms were first introduced by John Holland in 1960 and were extended by his student in 1989 [96]. After experiencing decades of development, it has been applied to many areas [97, 98, 99].

Genetic algorithms are mainly used to solve optimization and search problems [28]. An abundance of research has shown the power of genetic algorithms to solve complex optimization problems. Gong and Yang [100] applied a genetic algorithm for an image processing study. A genetic algorithm was used to optimize the compatibility between corresponding points and the continuity of the disparity map by applying it using the quadtree structure in their study. The results show the algorithm generates better disparity maps than iterative-based cooperative algorithms and the SEA algorithm.

Srivastava *et al* [101] used a genetic algorithm to optimize software testing efficiency. They applied a genetic algorithm to cluster, in order to find the most critical path in a program. The results show the algorithm is better than the exhaustive search and local search techniques.

Genetic algorithms have also proved to be a good method to perform variable selection [102]. Tan *et al* [103] used a genetic algorithm for attribute selection in data mining. They proposed a GA-SVM hybrid method that implements GA to search for the best attribute set to produce a good classification performance. The results showed a big improvement after applying the GA-SVM hybrid method. Sikora *et al* [104] used a genetic algorithm to do sample and feature selection in their study. The results show a huge improvement by using a genetic algorithm to solve data mining problems.

In this thesis, a genetic algorithm is applied to a genome-wide association study to find the association between the human genome and colorectal cancer. The main method that is used in this thesis is an extension of standard genetic algorithms. We use double chromosomes instead of a single chromosome. A profusion of research has proved that multiple chromosome genetic algorithms can produce accurate results [105, 106, 107]. Cavill *et al* [108] introduced a genetic algorithm that could perform simultaneous variable and sample selection. In their genetic algorithm method, they used two fixed lengths binary strings as the chromosome to represent sample selection and feature selection. They implemented this method in a metabonomics study to predict the toxicity of the liver and kidney. The average accuracy of this method was 64.52%. Additionally, they found some metabolites that are associated with hepatotoxicity or nephrotoxicity. However, this method has a good potential for

improvement. The behaviour methods are very simple. All the behaviours methods this study used are the most fundamental methods. The parameters of this genetic algorithm method were not selected by a benchmark. This means the performance still had a big climbing space. We upgraded this method by using different behaviour methods, including but not limited to selection function, mutation function, and crossover function. A better performance for the old mechanism can be expected by upgrading the behaviour functions.

2.6 Summary

In this chapter, we first discussed colorectal cancer and its risk. We then introduced the concept, the background, and the challenge of genome-wide association studies. We showed the background of machine learning and listed some GWAS that used machine learning methods. Then we described the evolutionary computing technique. We showed the related works of evolutionary computing, and the usual design of an evolutionary algorithm. Finally, we introduced genetic algorithms and demonstrated some of its related works. It enhanced many research areas to gain new insights. Genetic algorithms have shown a great ability to solve complex problems. However, only a small part of GWAS studies have applied genetic algorithms. The performance of genetic algorithms on GWAS is still unclear and needs to be tested. The connection of GWAS and genetic algorithms is worth pursing for its good potential.

Chapter 3

Data and Methods

3.1 Data Processing

The data we used in this research are genetic variants in diseased cases and healthy controls. The colorectal cancer GWAS case-control dataset was collected from Newfoundland and Labrador, Canada. Two datasets were obtained from the Newfoundland Familial Colorectal Cancer Registries (NFCCR). These two dataset include 265,195 SNPs in total. The participants contain 656 cases and 496 controls. The cases in this data set were diagnosed from 1999 to 2003. All the participants were 20 to 74 years old. The details of the original data are shown in Table 3.1.

To make sure the dataset is suitable for the machine learning analysis, data preprocessing needs to be done. Plink [109] is a whole-genome association analysis toolset that is designed flexibly to perform a wide range of basic, large-scale genetic analyses. It is a free, open-source tool to perform genetic data processing. It is widely used in GWAS for data management, basic statistics, and linkage disequilibrium (LD)

	Dataset1	Dataset2
SNPs	1236084	1134514
Individuals	696	656
Males	418	393
Females	278	263
Total	696	656
Cases	200	656
Controls	496	0

Table 3.1: Data description.

calculation, etc.

Firstly, we used Plink to merge two datasets based on the SNPs that they both have. The SNPs of these two datasets are not completely the same. We want as many samples as possible to have a better performance. Hence, we need to merge these two datasets before the computational investigation.

3.1.1 Quality control

We used Plink to apply quality control [110]. As well as applying individual quality control, we also conduct marker quality control to maximize the number of markers that remain in the merged dataset [110].

3.1.1.1 Individuals' quality control

The goal of this step is to remove individuals with low quality data from the dataset. Firstly, individuals with discordant sex information need to be deleted. We calculated

the mean homozygosity across X chromosome markers for each individual to remove individuals with discordant sex data.

The sex chromosome is not useful for data selection because it is an aneuploidies and genotyping artifact [111]. Sex chromosomes may lead to ambiguous results, we remove the sex chromosome from the dataset to optimize the performance.

Some samples in this dataset are not fully filled. Individuals with a high missing data rate are identified. Also, heterozygosity plays a key role in the quality of the dataset. High heterozygosity denotes many genetic variabilities. Low heterozygosity means little genetic variability. Too high or too low heterozygosity is not normal. Individuals with unusual heterozygosity should be removed. We calculate the heterozygosity value for every individual and remove the individuals beyond $\pm SD$. The formula that calculates heterozygosity is:

$$PM = \frac{N(NM) - O(HOM))}{N(NM)} \tag{3.1}$$

where $N(NM)$ is the number of non-missing genotypes, and $O(HOM)$ is the number of homozygous genotypes.

3.1.1.2 Makers' quality control

The goal of the makers' quality control is to remove substandard SNPs. Some SNPs have a high missing rate. These SNPs could cause difficulty for later work. Hence, we removed those SNPs with a missing rate higher than 0.05.

Checking Hardy-Weinberg Equilibrium (HWE) is an important step in the quality control analysis of markers in GWAS data [112]. In the Hardy-Weinberg theory, allele and genotype frequencies are predictable from generation to generation. The bias

of Hardy-Weinberg equilibrium could mean potential genotyping errors, population stratification, or even actual association to the trait under study [113]. Those SNPs that are greatly associated with the disease and also show highly significant departures from HWE, especially in controls, should be analyzed. We checked the deviation from HWE in the controls and generated a p-value for HWE's deviation for every SNP. Then SNPs with an HWE greater than 1×10^{-4} were removed.

Minor allele frequency (MAF) is also an important metric to filter SNPs for quality control. The statistical power of rare SNPs is deficient. SNPs with extremely low minor allele frequency should be removed from the dataset to obtain a better performance. We calculate the MAF for every SNP and only keep SNPs with an MAF greater than 0.05.

3.1.1.3 Linkage disequilibrium pruning

We did Linkage Disequilibrium (LD) pruning to delete those variables that are duplicated or related to others. Because of the genetic diversity of the samples and the density of SNPs, considerable redundancy could exist in loci, meaning that plenty of SNP pairs may have an extremely high linkage disequilibrium [114]. It is recommended to remove SNPs based on high levels of pairwise LD [115]. We calculate the Pearson Correlation Coefficient for every pair of SNPs:

$$r_{ij}^2 = \frac{(p_{ij} - p_i \times p_j)^2}{(p_i - p_i^2) \times (p_j - p_j^2)} \tag{3.2}$$

where p_i,p_j are minor allele frequencies of i_{th} and j_{th} SNP, p_{ij} represent the frequency two-marker haplotypes. We remove one SNP from the correlated pair with r greater than 0.6, keeping the one with the largest minor allele frequency.

We also calculate the Identity by Descent (IBD) of all pairs of samples based on the reduced marker set. The value of IBD demonstrates the potential relationship between the two samples. The higher the IBD value, the closer relationship they have. We remove one sample from each pair with an IBD value greater than 0.25.

3.1.1.4 Imputation

After previous steps, some missing values can still be found in the dataset. To complete the dataset, regarding the rest of the missing SNPs, we filled it with the most frequent value of the corresponding SNPs [74].

3.1.2 Filter

After the preliminary processing, we had a dataset that contains 18,5180 SNPs and 1,098 individuals. To deal with a dataset that contains a huge amount of features would be extremely time-consuming. After the benchmark, we estimated it would take more than 20 days to finish our algorithm on this dataset. Hence, the dataset needed more processing. We decided to implement a data filter to reduce the number of SNPs to 1000.

3.1.2.1 ReliefF

Relief [116] is a well-known filter method. It calculates the weight for each feature based on feature value differences between nearest-neighbour instance pairs. It uses Euclidean distance to calculate the distance. Features with top-weight are selected to remain. An observed feature difference between the feature and a neighbouring instance that has the same class leads to a weight decrease, called a hit. However, an

observed feature difference between the feature and a neighbouring instance that has a different class leads to weight growth, called a miss.

Kononenko developed a new feature selection method, ReliefF, inspired by Relief [117]. Instead of using Euclidean distance, ReliefF uses a taxicab metric to calculate the distance. The distance calculation is :

$$dist(S_i, S_j) = \sum_{a \in A} diff(a, S_i, S_j) \tag{3.3}$$

where A means all SNPs. The difference(diff) calculation is:

$$diff(a, S_i, S_j) = \begin{cases} 0 & genetype(a, S_i) == genotype(a, S_j) \\ 1 & genetype(a, S_i)! = genotype(a, S_j) \end{cases} \tag{3.4}$$

Hence, an individual's nearest neighbour is the individual that has the most amount of SNPs that are the same genotype.

3.1.2.2 SURF and TURF

Spatially uniform ReliefF (SURF) is an extension to ReliefF developed by Greene *et al* [118]. They select neighbours within a constant distance as the nearest neighbours of the instance instead of choosing a fixed number of nearest neighbours.

The tuned ReliefF (TURF) algorithm [119] provides a way to improve the performance of ReliefF by running it multiple times. It is widely used to address noise in large datasets by doing recursive elimination of features and the iterative application of Relief. It is recommended to use an iterative Relief approach to deal with datasets with large size [120].

We decided to use a method that combined tuned reliefF (TURF) [119] and spatially uniform reliefF (SURF) [118] to filter the dataset. It has been proved that

```
1: while |F| > 1000 do
2:   for i = 1 to |S| do
3:     hit = s ∈ S | ph(s) == ph(S_i) and dist(s, S_i) < MD
4:     miss = s ∈ S | ph(s)! = ph(S_i) and dist(s, S_i) < MD
5:     for f in F do
6:       for h in hit do
7:         W[f] = w[f] − diff(f, R_i, h)/(|S| ∗ |hit|)
8:       end for
9:       for m in miss do
10:        W[f] = w[f] + diff(f, R_i, m)/(|S| ∗ |miss|)
11:      end for
12:    end for
13:  end for
14:  remove %1 SNPs with the least weight W from F
15: end while
```

Figure 3.1: The SURF + TURF algorithm

using the combination of TURF and SURF can produce an accurate performance [121]. The pseudocode of this combination is shown in Fig.3.1, where F and S denote the features (SNPs) and samples, MD is the mean distance of all samples, and ph is the phenotype.

After finshing the data process, 997 SNPs remain in the dataset. This includes 1,098 individuals, as the samples contain 626 cases and 472 controls. The dataset's details are shown in Table 3.2.

Variables	Amount
SNPs	997
Samples	1098
Case	626
Control	472

Table 3.2: Dataset after preprocessing and filtering

3.2 Methods

In this thesis, we use a genetic algorithm [88], a branch of the evolutionary algorithm, to achieve research objectives. The mechanisms of the evolutionary algorithm were inspired by biological evolution. This process includes reproduction, mutation, recombination, and selection. Each individual in an evolutionary algorithm is a potential solution to the problem. Furthermore, every individual's quality is determined by a fitness function. Then the evolution of the real biosphere is simulated. New generations are produced and those individuals with high quality are selected, until the best solution is found or the max number of generation is reached.

The genetic algorithm method we use in this thesis is an extended version of the traditional genetic algorithm method. Unlike the most common genetic algorithms method, which has only one chromosome, this method uses two chromosomes, which are two binary strings that represent sample and feature selection.

3.2.1 Classifier

To evaluate individuals' fitness, we need a suitable classifier for the fitness function. We tested 4 different classifiers, K-Nearest Neighbours algorithm (KNN) [122], Support Vector Machine (SVM) [123], random forests (RF) [124], and Gradient Boosting. K-Nearest Neighbours is one of the most popular statistical methods. KNN classifies the target by a plurality vote of its neighbours. Support Vector Machine is a supervised machine learning algorithm that is widely used to deal with classification problems and regression analysis. SVM solves problems by constructing a hyperplane that is defined by the largest distance to the nearest training-data point of any class. Random forests is a widely used ensemble learning method for classification, regression, and other tasks. It finishes tasks by building multiple decision trees and outputting the class that is the most common value of the classification. Gradient boosting is a technique that generates prediction models in the form of an ensemble of weak prediction models, like decision trees, for regression and classification problems.

We used the receiver operating characteristic (ROC) curve to compare the performance of different classifiers. The ROC curve shows the trade-off between the true positive rate and the false positive rate. Classifiers with curves that closer to the top-left corner indicate better performance. We recorded the accuracy and the ROC-AUC score of the four classifiers. The ROC-AUC score presents the area under the roc curve. Comparing their performance on the dataset, shown in Table 3.3, Fig.3.2, Fig.3.3, Fig.3.4, and Fig.3.5, the performance of SVM is better than that of other classifiers. Therefore, we decided to use SVM for computing the fitness function prediction accuracy.

Classifier	Accuracy
KNN	0.48521312805716477
SVM(C=1)	0.6932332499304976
RF(N=10)	0.5908632569183028
Gradient Boosting	0.6102618548490109

Table 3.3: Mean accuracy of classifiers

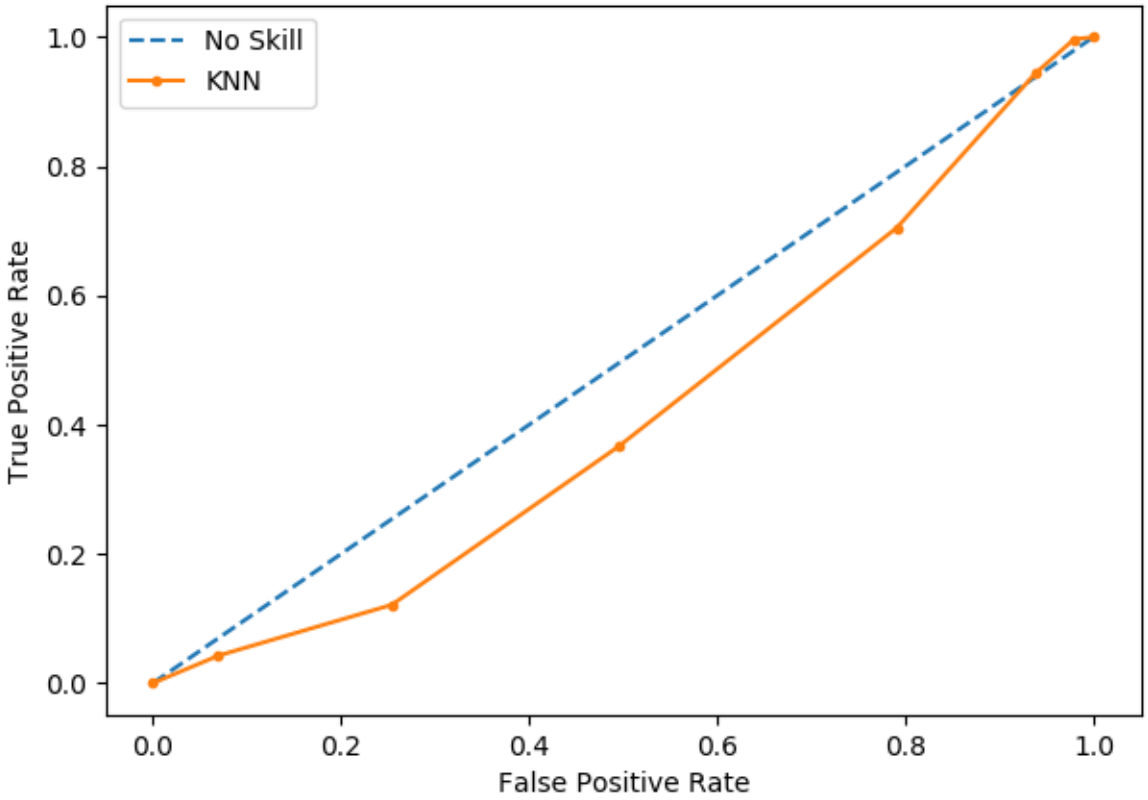


Figure 3.2: KNN ROC Curve.

This demonstrates of the performance of K-Nearest Neighbours on the original dataset.

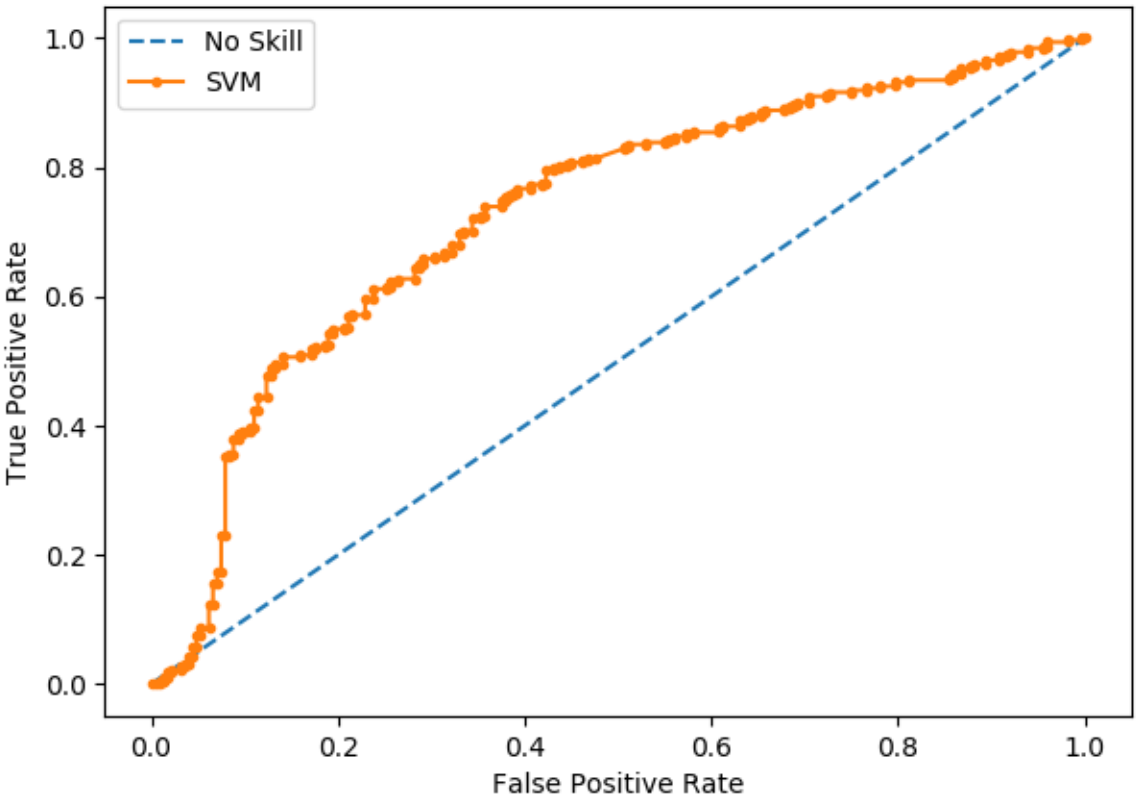


Figure 3.3: SVM ROC Curve.

This demonstrates of the performance of Support Vector Machine on the original dataset.

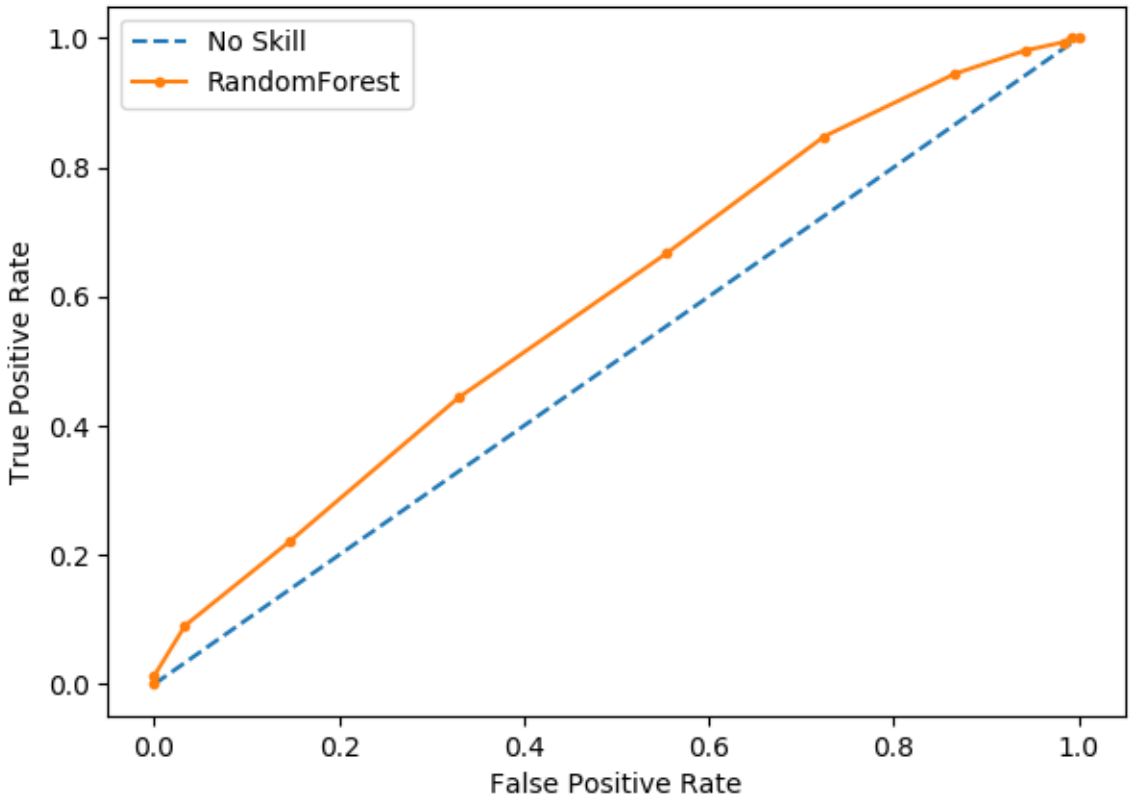


Figure 3.4: RF ROC Curve.

This demonstrates of the performance of random forests on the original dataset.

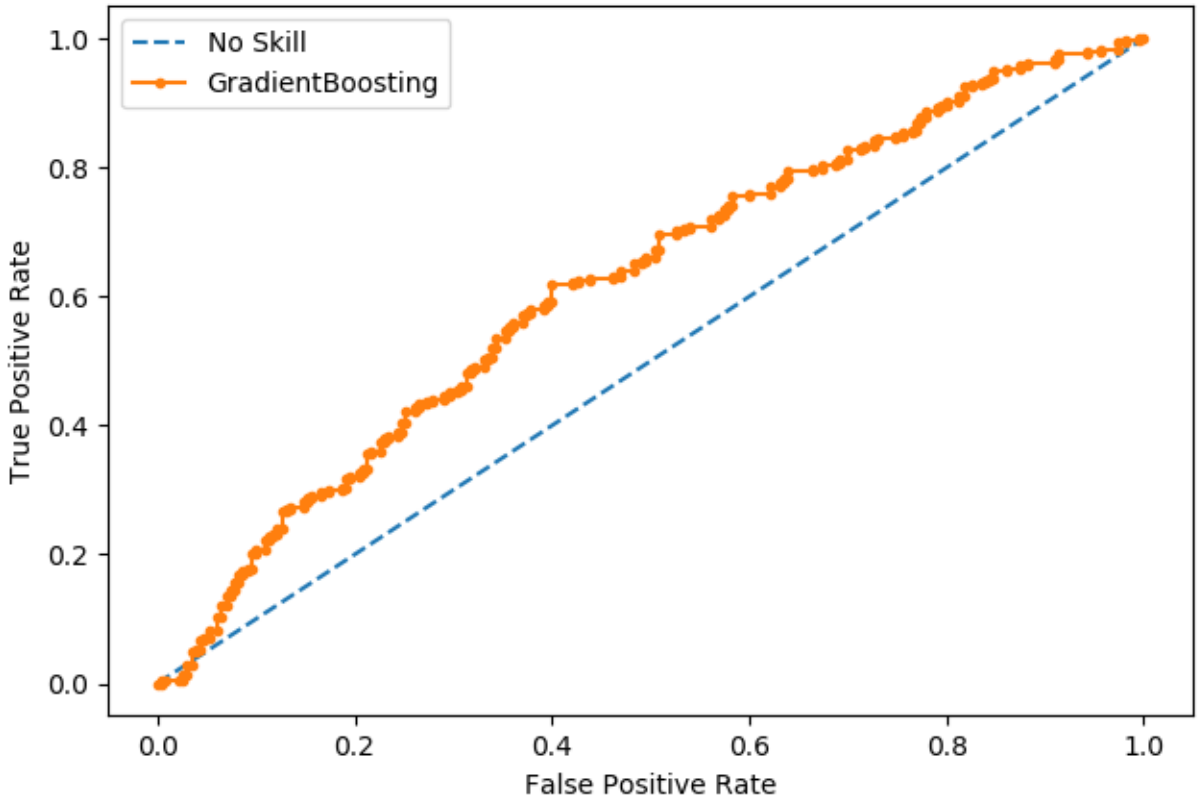


Figure 3.5: GB ROC Curve.

This demonstrates of the performance of gradient boosting on the original dataset.

3.2.2 Parameters tuning

To have excellent GA performance, finding a set of the most suitable parameters is essential. We performed parameter tuning of our GA. The parameters of GA normally include crossover probability and mutation probability. Population size is an essential factor that could affect the performance of GA. Because we use tournament selection as the parents selection in this GA method, the size of the tournament is also important. Hence, all those four parameters need to be optimized in parameter tuning.

We extracted a small subset that contains 100 features and 100 samples from the dataset for parameters tuning. All features were randomly selected. The numbers of cases and controls in the subset were equal. We tested 18 different combinations of mutation rate, crossover rate, and tournament size. Each combination was run 10 times on the small subset. Then we collected each combination's average accuracy. By comparing their performance, the best parameter combination is 0.3 for mutation rate, 0.7 for crossover rate, and 15 for tournament size. The results of parameter tuning are shown in Table 3.4.

For population size, we tested two different values, 100 and 1000, on the small data subset. The other parameter values were 0.3 for mutation rate, 0.7 for crossover rate, and 15 for tournament size. The results show that the GA with 100 individuals in the population have better performance, 0.68, compared to the accuracy of GA with 1000 population size, 0.65. We decided to set the population size to 100.

Mutation rate	Crossover rate	Tournament size	Accuracy
0.3	0.3	5	0.598718258
0.3	0.3	10	0.60094505
0.3	0.3	15	0.615822117
0.3	0.5	5	0.634188042
0.3	0.5	10	0.637262879
0.3	0.5	15	0.618997698
0.3	0.7	5	0.707411643
0.3	0.7	10	0.718067985
0.3	0.7	15	0.720809128
0.5	0.3	5	0.625417945
0.5	0.3	10	0.623152186
0.5	0.3	15	0.621603929
0.5	0.5	5	0.716424847
0.5	0.5	10	0.704259803
0.5	0.5	15	0.697076913
0.7	0.3	5	0.705606232
0.7	0.3	10	0.685080314
0.7	0.3	15	0.697148618

Table 3.4: Parameter tuning

Parameters	*Setting*
Representation	Double binary Strings
Mutation	Bit flip
Mutation rate	0.3
Crossover method	Two-point crossover
Crossover rate	0.7
Parent selection	Tournament selection
Tournament size	15
Survivor selection	Fitness-based replacement
Fitness function	SVM accuracy

Table 3.5: Parameters setting.

3.2.3 Proposed GA

After the parameter tuning, the final setting of our GA is shown in Table 3.5. The process is shown in Fig.3.6.

3.2.3.1 Representation

In this study, we apply a double chromosome representation of the genetic algorithm. The two chromosomes represent the result of sample selection and feature selection as the phenotype. Every bit of a binary string demonstrates if the corresponding feature or sample is selected, where 1 is positive and 0 is negative. Each individual represents a candidate selection of samples and features.

3.2.3.2 Population initialization

First of all, generate the population which contains 100 individuals that hold possible solutions. The chromosomes of individuals are randomly generated. Every bit of chromosome has a half chance to be 1 or 0. After the whole population is generated, evaluate all individuals' fitness to prepare for the evolution process.

3.2.3.3 Fitness function

Because the number of the cases is much bigger than the controls in the dataset, using regular accuracy as the fitness of individuals would let case samples dominate the sample selection result. To avoid that, we used balanced accuracy in this method. Unlike the regular accuracy calculation that only considers true positive predictions, balanced accuracy considers not only true positive predictions but also involves true negative predictions. The calculation of balanced accuracy is shown below:

$$Balanced\ accuracy = (\frac{TP}{TP+FP} + \frac{TN}{TN+FP})/2 \tag{3.5}$$

where TP, FP, TN, and FN denote true positive, false positive, true negative, and false negative.

Additionally, we adopted ten-fold nested cross-validation to define the fitness of every individual in the population. Each time an individual is being evaluated, randomly separate the samples, which are selected by the corresponding individual's sample selection phenotype, into ten partitions, and keep the features that are chosen by the individual's feature selection phenotype. Then, apply SVM ten times to determine the balanced accuracy. In each of the ten determinations, the SVM classifier using nine partitions and the classifier is evaluated on the remaining testing partition.

The mean balanced accuracy that is generated by SVM is the fitness of the individual.

3.2.3.4 Parent selection

Next, the evolution process begins. To allow better individuals to become parents of the next generation, high-quality individuals should be given more opportunities to become parents. Nevertheless, low-quality individuals in the population should have a small, but positive chance to be parents. We use tournament selection to select individuals as the parents to generate new individuals. 15 individuals are randomly selected from the population to become a group. Then the individuals with the best fitness in the group are chosen to be the parents, in order to produce new potential solutions for the optimization problem.

3.2.3.5 Crossover and mutation

To create new individuals from old ones, the population needs variation to generate a new potential solution. Crossover and mutation are the main variation operators in evolutionary algorithms.

In order to merge information from two parents with high fitness genotypes, offspring that combines both of those desirable features should be produced. Two-point crossover is one of the most popular crossover methods used in genetic algorithms. It randomly picks two points on the genotype of parents individuals and switches the middle part between these two points to generate offspring.

If an evolutionary algorithm only performs crossover during the evolution process, it could be easily trapped in a locally optimal solution. To avoid this, the population needs some random, unbiased changes. In this study, the mutation method is bit

1: Initialize the population
2: G=0
3: **while** $G < 1500$ **do**
4: Evaluate all individuals in the population
5: Select 10 individuals to generate 10 offspring
6: Mutation
7: Delete 10 individuals with the least fitness
8: $G = G + 1$
9: **end while**

Figure 3.6: The evolution process

flipping. When a mutation occurs, every bit of the individual has a 50% chance to changed from 1 to 0 or 0 to 1.

3.2.3.6 Survivor selection

Unlike parent selection, which is stochastic, survivor selection is deterministic. Since the population size is fixed to 100 and we want to favour the individuals with higher fitness, the next step is eliminating those individuals with the worst fitness to keep the quality of the population as advantageous as possible. The same number of individuals as the amount of newly created offspring are deleted.

The end of the survivor selection is the end of one generation. After generating the high fitness parent's new offspring and removing low fitness individuals, the quality of the population will be improved. Then this evolution process is repeated until the genetic algorithm reaches the max number of generations.

3.3 Results visualization

In order to obtain more details from the working of the genetic algorithm, we use different visualizations in order to have a better view of the results. Data visualization is an important part of the research field. Using tables, charts, or images to summarize and present the results is very popular as if is very helpful to gain insight.

We illustrate the performance of GA. To demonstrate the quality of our method, we show some metrics of the individuals in the population throughout the run of the genetic algorithm and compare these with other methods. Fitness metrics is widely used to evaluate the performance of genetic algorithm methods, and especially the best individual's fitness in the population. The line graph that contains the mean values of best fitness, mean fitness and worst fitness of 100 runs shows the improvement of every individual during the evolution. It signifies whether the potential solutions evolve or not during the time of the process. Moreover, we compare the classifiers' performance in the dataset before the genetic algorithm method and the dataset after using the GA method to show the data selection's work.

To gain more insight from the results of sample and feature selection, we do importance analysis. We use figures to show samples and features that have a high frequency. Over-selected variables mean they have shown that they play key roles in the dataset, compared to others. Also, we calculate the p-value for every feature in order to identify important SNPs. The p-value calculation is:

$$P(x) = \sum_{i=x}^{n} C_n^i (\frac{m}{i})^i (1 - \frac{m}{v})^{n-i} \tag{3.6}$$

where x means being selected x times. n is the number of runs, v is the number of total variables, and m is the variable's mean number that is selected.

A Q–Q (quantile-quantile) plot is a probability plot, which is a visualization method for comparing two probability distributions by plotting their quantiles against each other [90]. The Q-Q plot shows the difference between the distribution of the results and the standard normal distribution.

The sample selection results' visualization aims to show which samples have been over-selected. Additionally, we illustrate relationships between samples to identify which pair of cases are potentially the same sub-type of colorectal cancer.

Chapter 4

Results

After all of the parameters were set, we implemented our GA using DEAP [125]. We ran our GA method 100 times and collected every run's outcomes.

4.1 GA performance

To show our GA method's performance, we record the best, the worst, the mean fitness, and the standard deviation of the population for every 100 generations in each run. Fig.4.1 shows the fitness changes during evolution. From the consistent growth of all three fitness metrics and the large drop of standard deviation, it can be seen that the mean accuracy of the population keeps growing during this process.

4.2 Comparison

Comparing the genetic algorithm method with only sample selection or feature selection, the double chromosomes method that applies both selections at the same

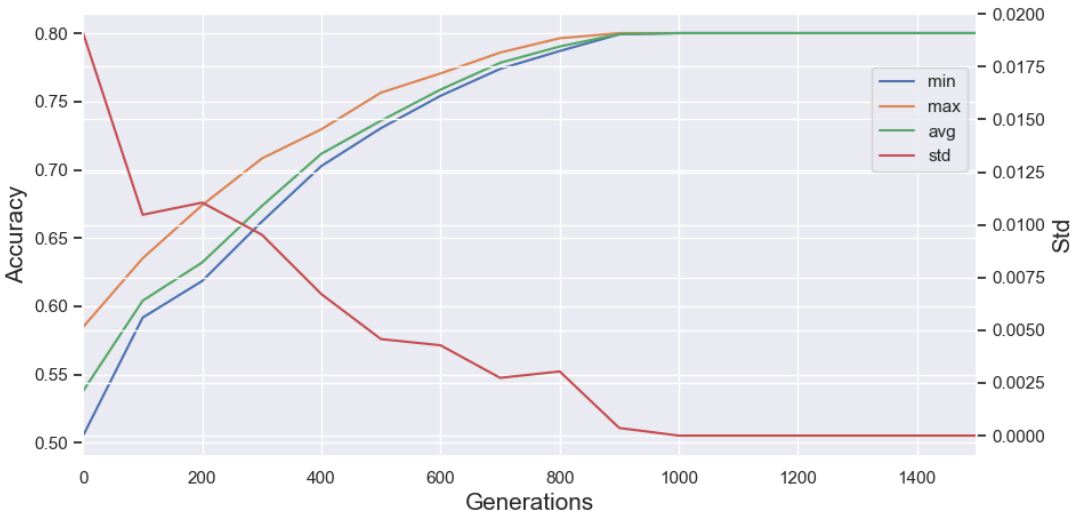


Figure 4.1: Mean accuracy and standard deviation during evolution over 100 runs

time has better performance than the others. In order to get a good comparison, we computed several metrics. Recall, also known as sensitivity, is the fraction of the total amount of relevant instances that have actually been retrieved. The calculation of recall is:

$$Recall = \frac{True\ positive}{True\ positive + False\ negative} \tag{4.1}$$

We also used the ROC-AUC score. All the metrics were calculated by Scikit-learn [126].

Fig.4.2 shows the metrics of different genetic algorithms. We compared our simultaneous GA method to standard GA methods using only the variables or samples selection. The simultaneous method shows better performance than other two method in all three metrics.

The best data selection results also provide an improvement to classifiers. After sample selection and feature selection by this approach, the balanced accuracy of

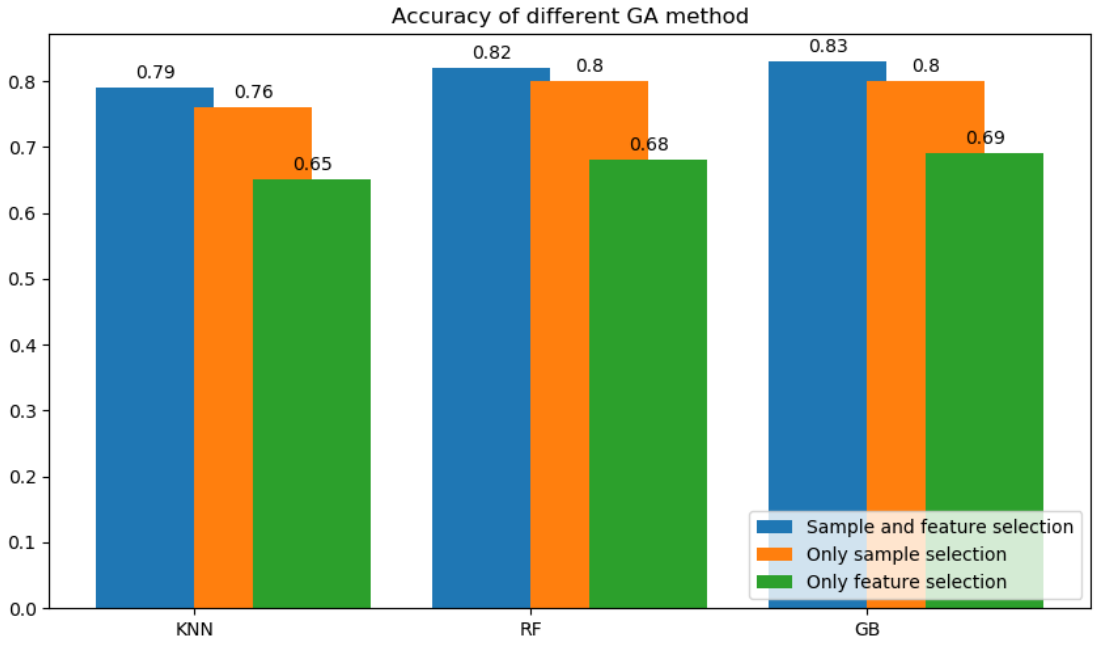


Figure 4.2: Performance for the different GA methods.

Comparison between the GA method with only sample selection, only feature selection, and both selections.

SVM has a significant increase, from 69% to 85%. The average balanced accuracy of the 100 runs is 79%. The distribution of all results is shown in Fig.4.3. All the results demonstrate good performance for prediction. Moreover, as it showed in Fig.4.4, other classifiers' performance also improves. Along with the SVM classifier, other classification methods, KNN, RF, and GB also have an increase of accuracy.

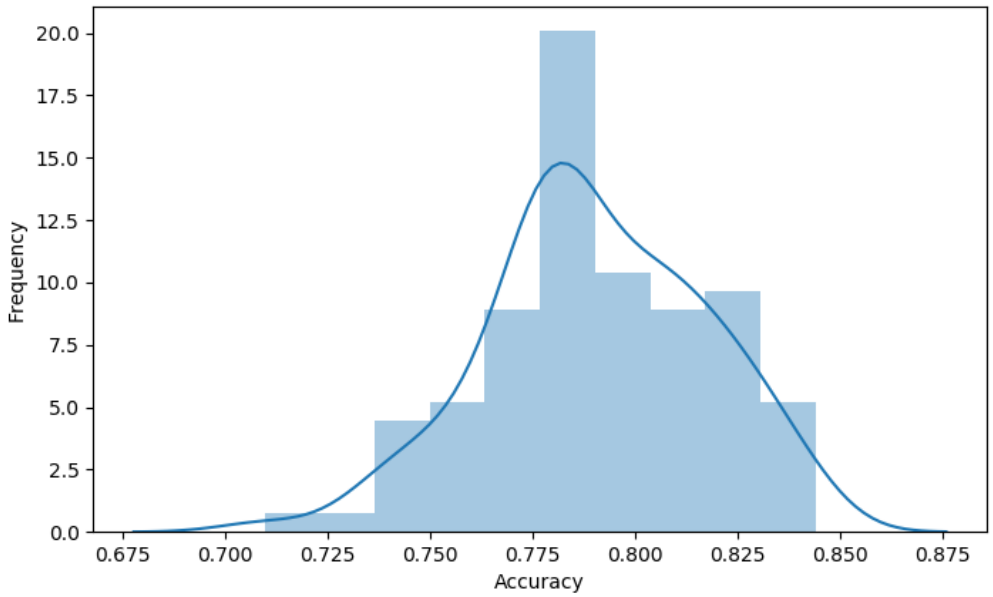


Figure 4.3: The distribution of results over 100 runs.

Every best result from the 100 runs shows good performance. The most frequent accuracy over the 100 runs is around 78%. The worst accuracy is 72.5%, and the best is 85%.

4.3 Importance analysis

4.3.1 Features analysis

We recorded the frequency of every feature occurring in the best evolved predictive models of the 100 GA runs. Based on the results that we collected, we did data visualization to enable a better view.

Fig.4.5 illustrates the result of SNPs' importance assessment. As the figure shows, some SNPs played dominant roles in the results of 100 runs. Some of them occurred 90 times. This result denotes that these over-selected SNPs have a high possibility to

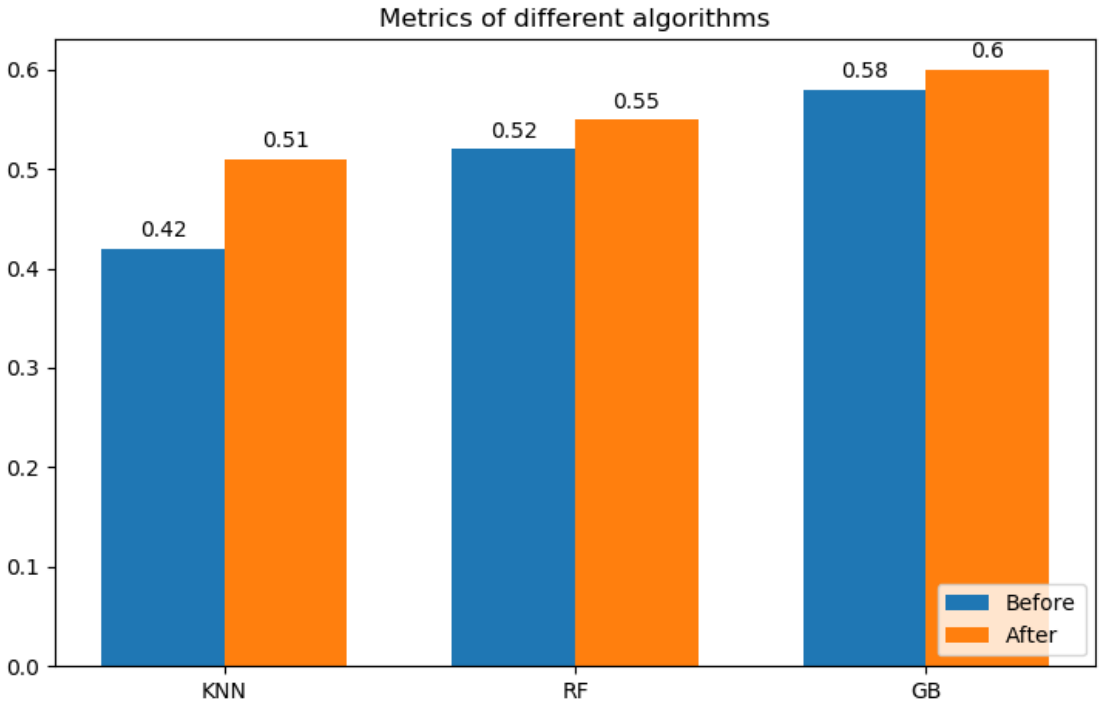


Figure 4.4: The performance of other classifiers before and after the data selection Comparison of the accuracy of three different classifiers using the original data and the data after the selection of the GA method.

have associations with colorectal cancer. The distribution of occurrence of all SNPs is shown in Fig.4.7. Moreover, we calculated the p-values of all SNPs in the dataset. The p-values of the top over-selected SNPs are shown in Table 4.1. Based on the p-values, we made a Q–Q (quantile-quantile) plot, which is Fig.4.6 . The distribution of the occurrence and the Q-Q plot of all SNPs provide extra proof that those SNPs are important for colorectal cancer. Table 4.2 shows the information of top 22 frequent SNPs. The distribution and p-values give more confidence that these top SNPs are relevant to the diseases.

The top over-selected genes show relevance with colorectal cancer. rs8015314 has a

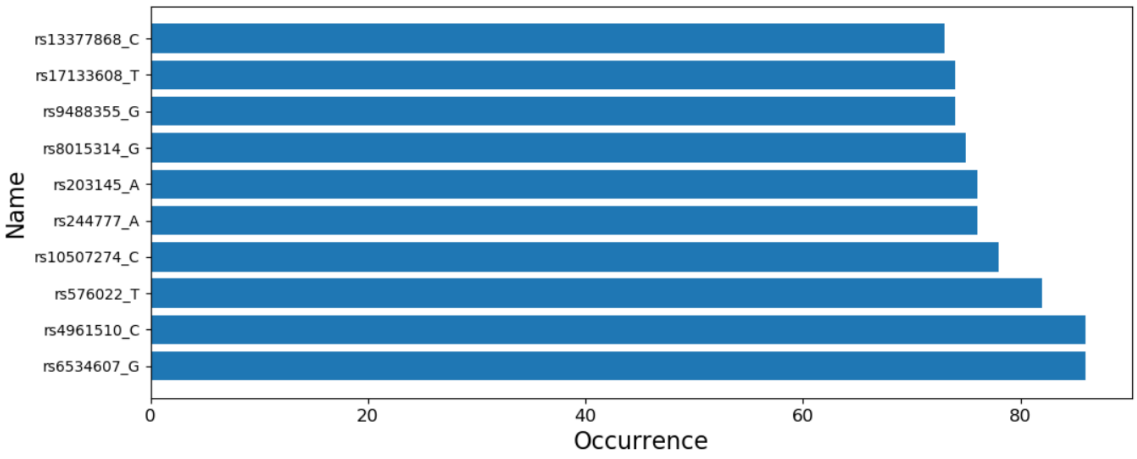


Figure 4.5: Feature importance results.

The Top 10 features that were over-selected over 100 runs.

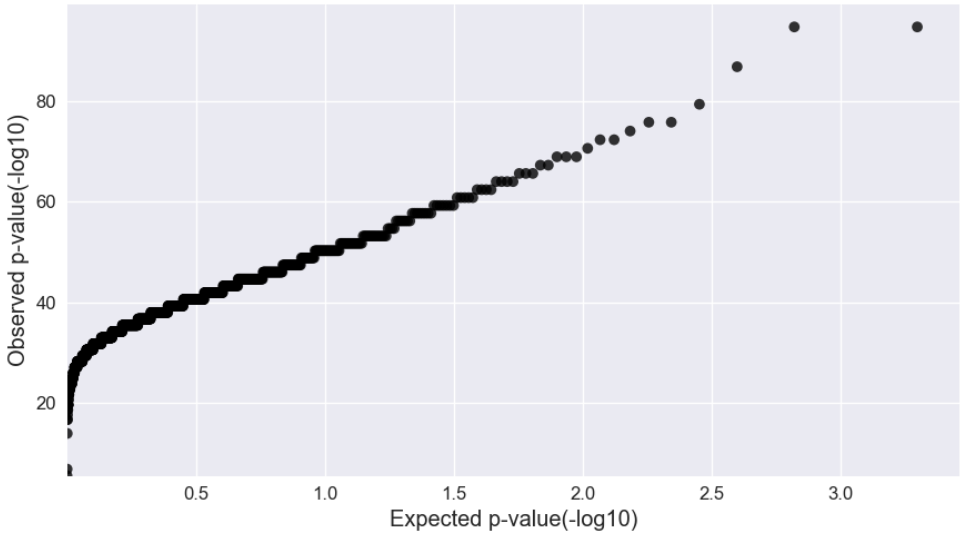


Figure 4.6: Q-Q plot.

P values for each SNP are sorted from largest to smallest. P values that move towards the y-axis means they are more significant than expected under the null hypothesis.

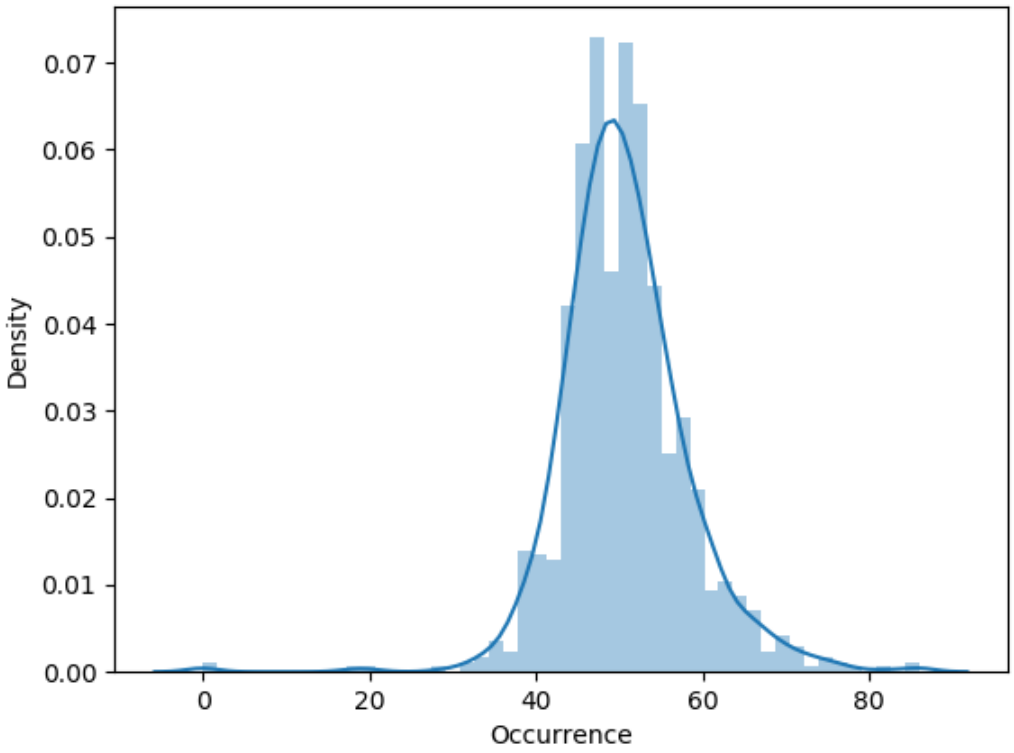


Figure 4.7: The distribution of occurrence of SNPs over 100 runs.

The distribution shows that the number of occurrence of most SNPs is around 50.

high gene expression correlation with a colon-sigmoid tissue-specific gene, LINC02279. NEU3 may have an effect in modulating the ganglioside content of the lipid bilayer. It has been widely proven that is associated with colorectal cancer [127, 128, 129]. Many research has found that RMST plays a role in cancer diseases [130, 131]. Both rs2041396 and rs7799059 have high gene expression correlations with colon-transverse tissue-specific genes. Long Non-coding RNAs (LnRNA) have been proven by many research that it plays an important role in cancer [132]. Tab2 is involved in heart development and has been identified that is relevant to breast cancer [133].

4.3.2 Samples analysis

We did the same work for the sample selection results. Fig.4.8 illustrates samples with the most occurrences. The bars with red colour denote the cases, and the blue bars are the controls. Because we used balanced accuracy instead of regular accuracy, the selection results were not dominated by the samples from any class. Also, some samples were selected multiple times through 100 runs. This means the genotype of these individuals has high associations with colorectal cancer.

We selected the top 100 over-selected case samples and checked their relationships with each other. We recorded the number of simultaneous occurrences of each pair of samples. Fig.4.9 is the heat-map that shows the relationships between case samples based on the number of occurrences. It clearly shows that some cases have high associations with others, which denotes that they are very likely to belong to the same subtype of colorectal cancer.

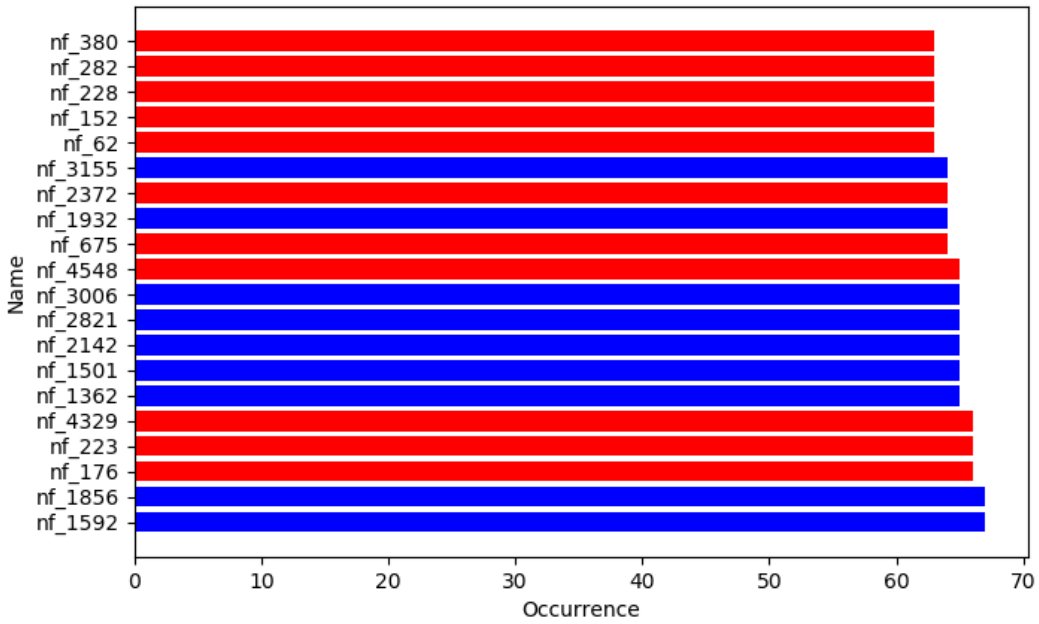


Figure 4.8: Samples importance results.

The top 20 over-selected samples over 100 runs

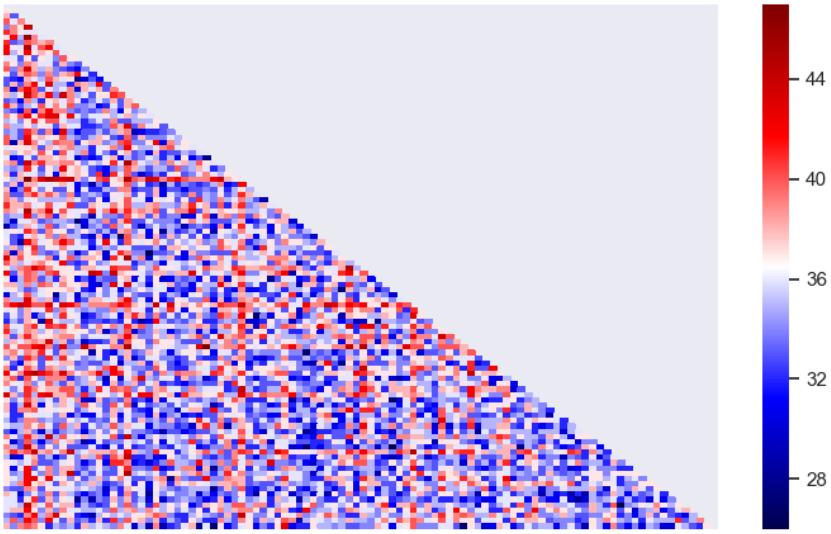


Figure 4.9: The association between samples.

Each point in this figure represent a association of two samples. The more simultaneous occurrences they have, the redder their point is.

4.4 Summary

In this section, we first showed the performance of our genetic algorithm method by demonstrating the evolution progress in Section 4.1. Then we showed the improvement of classical classifiers by using our method in Section 4.2 and compared our method to other GA methods. Finally, we performed an importance analysis in Section 4.3 to show the associated SNPs and individuals we found and showed their relation. The results showed that our GA method was able to achieve great performance and can find associated SNPs in GWAS datasets.

Name	Count	P-value
rs6534607_G	86	1.96E-95
rs4961510_C	86	1.96E-95
rs576022_T	82	1.62E-87
rs10507274_C	78	4.60E-80
rs244777_A	76	1.72E-76
rs203145_A	76	1.72E-76
rs8015314_G	75	9.73E-75
rs9488355_G	74	5.21E-73
rs17133608_T	74	5.21E-73
rs13377868_C	73	2.65E-71
rs17042892_T	72	1.28E-69
rs10254969_A	72	1.28E-69
rs4795690_A	72	1.28E-69
rs17799628_C	71	5.92E-68
rs2041396_C	71	5.92E-68
rs7799059_T	70	2.60E-66
rs9634692_C	70	2.60E-66
rs2206451_C	70	2.60E-66
rs9827966_A	69	1.09E-64

Table 4.1: The p-values of over-selected features

SNP	Position	Gene
rs8015314_G	chr14:94972891	None
rs17133608_T	chr11:75012631	NEU3
rs13377868_C	chr12:83329519	None
rs9320356_C	chr6:111073218	None
rs1579244_C	chr12:97452764	RMST
rs244777_A	chr5:35366153	None
rs17799628_C	chr9:103680714	None
rs2041396_C	chr17:65423898	None
rs7799059_T	chr7:155998692	None
rs9634692_C	chr13:54353623	None
rs2206451_C	chr20:52775930	None
rs6534607_G	chr4:75106366	None
rs4961510_C	chr9:16962420	LncRNA
rs576022_T	chr6:149315830	TAB2
rs10507274_C	chr12:116723171	C12orf49
rs203145_A	chr6:138294011	ARFGEF3
rs9488355_G	chr6:114258132	HS3ST5/HDAC2-AS2
rs17042892_T	chr2:21699337	LINC01822
rs10254969_A	chr7:30067179	PLEKHA8
rs4795690_A	chr17:32298186	RHBDL3
rs9827966_A	chr3:8854430	LOC107984112
rs7190644_G	chr16:19582848	VPS35L

Table 4.2: Information of top SNPs.

name	Count	P-value
nf_1362	65	2.14E-58
nf_1501	65	2.14E-58
nf_2142	65	2.14E-58
nf_2821	65	2.14E-58
nf_3006	65	2.14E-58
nf_4548	65	2.14E-58
nf_675	64	7.18E-57
nf_1932	64	7.18E-57
nf_2372	64	7.18E-57
nf_3155	64	7.18E-57
nf_62	63	2.31E-55
nf_152	63	2.31E-55
nf_228	63	2.31E-55
nf_282	63	2.31E-55
nf_380	63	2.31E-55
nf_778	63	2.31E-55
nf_1190	63	2.31E-55
nf_1647	63	2.31E-55
nf_1651	63	2.31E-55

Table 4.3: P-values of over-selected samples.

Chapter 5

Conclusion

5.1 Summary

In this thesis, we proposed a new genetic algorithm method. We used this method to prioritize SNPs in the human genome in order to gain new insights into colorectal cancer. In this method, we first conducted data processing to generate a more efficient dataset for genetic algorithms. Then we performed grid search to find the best parameter setting to obtain the best results. Thirdly we applied this new genetic algorithm to the dataset. After we acquired the results, we did a results analysis for a better demonstration of the results.

The best result had an accuracy of 85% using SVM. The accuracy improved from 69% to 85%. Also, it is easy to observe that some SNPs play essential roles in colorectal cancer. We listed the SNPs and samples that dominated the results, which were generated by 100 runs. Moreover, we proved that the genetic algorithm can be a powerful tool to help researchers perform genome-wide association studies. By

analyzing the sample selection results, we can glimpse colorectal cancer's subtype and know which cases belong to the same subtype.

We contribute to the understanding of colorectal cancer, GWAS, and evolutionary computing. We developed a genetic algorithm method with double chromosomes that is able to achieve good performance in GWAS. A grid search was performed to find the best parameters to optimize the performance of GA. The results showed a significant improvement in classification accuracy compared to other methods. Our method found SNPs that were identified to have had important roles in colorectal cancer. The promising outcomes proved that evolutionary computing is a useful tool to help researchers perform GWAS.

5.2 Future work

This research still has much potential that can be explored. Our goal is not only to find associations between the human genome and colorectal cancer but also to identify subtypes of colorectal cancer. Many studies have proved that subtypes exist in colorectal caner [34, 56]. We want to correctly group cases to the correct subtype and find the most relevant SNPs for each subtype of colorectal cancer. We believe that with some upgrading to our method and better analysis, this can be achieved.

Because of the quality of the original dataset, there are a few false variables added during the data process, because some variables are filled with the most frequent value in the corresponding feature, which could slightly influence the performance of the method. In the future, we will apply this method to other datasets.

Bibliography

[1] Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

[2] Ebony B Bookman, Kimberly McAllister, Elizabeth Gillanders, Kay Wanke, David Balshaw, Joni Rutter, Jill Reedy, Daniel Shaughnessy, Tanya Agurs-Collins, Dina Paltoo, et al. Gene-environment interplay in common complex diseases: forging an integrative model—recommendations from an NIH workshop. *Genetic Epidemiology*, 35(4):217–225, 2011.

[3] Anthony P Polednak. Do physicians discuss genetic testing with family-history-positive breast cancer patients? *Connecticut medicine*, 62(1):3–7, 1998.

[4] Neil Risch and Kathleen Merikangas. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517, 1996.

[5] Jayakumar Ananthan, Alfred L Goldberg, and Richard Voellmy. Abnormal proteins serve as eukaryotic stress signals and trigger the activation of heat shock genes. *Science*, 232(4749):522–524, 1986.

[6] Akitada Ichinose, Erik S Espling, Junki Takamatsu, Hidehiko Saito, Koichi Shinmyozu, Ikuro Maruyama, Torben E Petersen, and Earl W Davie. Two types of abnormal genes for plasminogen in families with a predisposition for thrombosis. *Proceedings of the National Academy of Sciences*, 88(1):115–119, 1991.

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

[8] David JA Goldsmith and Adrian Covic. Coronary artery disease in uremia: Etiology, diagnosis, and therapy. *Kidney international*, 60(6):2059–2078, 2001.

[9] Roy C Page. The etiology and pathogenesis of periodontitis. *Compendium of continuing education in dentistry (Jamesburg, NJ: 1995)*, 23(5 Suppl):11–14, 2002.

[10] Tracy Murray-Stewart, Yanlin Wang, Andrew Goodwin, Amy Hacker, Alan Meeker, and Robert A Casero Jr. Nuclear localization of human spermine oxidase isoforms–possible implications in drug response and disease etiology. *The FEBS journal*, 275(11):2795–2806, 2008.

[11] Alan Wright and Nicholas Hastie. *Genes and Common Diseases: Genetics in Modern Medicine*. Cambridge University Press, 2007.

[12] Anne M Glazier, Joseph H Nadeau, and Timothy J Aitman. Finding genes that underlie complex traits. *Science*, 298(5602):2345–2349, 2002.

[13] What are Genome Wide Association Studies (GWAS)? https://www.ebi.ac.uk/training-beta/online/courses/gwas-catalogue-exploring-snp-trait-associations/what-is-gwas-catalog/what-are-genome-wide-association-studies-gwas/. Accessed: 2020-09-30.

[14] Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre. Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, 20(8):467–484, 2019.

[15] Julie Bogaert and Hans Prenen. Molecular genetics of colorectal cancer. *Annals of gastroenterology*, 27(1):9, 2014.

[16] Jiliang Tang, Salem Alelyani, and Huan Liu. Feature selection for classification: A review. *Data classification: Algorithms and applications*, page 37, 2014.

[17] Mike Oaksford and Nick Chater. Optimal data selection: Revision, review, and reevaluation. *Psychonomic Bulletin & Review*, 10(2):289–318, 2003.

[18] Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[19] Junfei Qiu, Qihui Wu, Guoru Ding, Yuhua Xu, and Shuo Feng. A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing*, 2016(1):1–16, 2016.

[20] Ting Hu, Yuanzhu Chen, Jeff W Kiralis, Ryan L Collins, Christian Wejse, Giorgio Sirugo, Scott M Williams, and Jason 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.

[21] Ting Hu, Marco Tomassini, and Wolfgang Banzhaf. Complex network analysis of a genetic programming phenotype network. In *European Conference on Genetic Programming*, pages 49–63. Springer, 2019.

[22] Thomas Bäck, David B Fogel, and Zbigniew Michalewicz. *Handbook of Evolutionary Computation*. CRC Press, 1997.

[23] L Davis, editor. *Handbook of Genetic Algorithms*. Van Nostrand Reinhold, 1991.

[24] Ágoston E Eiben, Robert Hinterding, and Zbigniew Michalewicz. Parameter control in evolutionary algorithms. *IEEE Transactions on Evolutionary Computation*, 3(2):124–141, 1999.

[25] Gary B Fogel and David W Corne. *Evolutionary Computation in Bioinformatics*. Elsevier, 2002.

[26] Sankar K Pal, Sanghamitra Bandyopadhyay, and Shubhra Sankar Ray. Evolutionary computation in bioinformatics: A review. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 36(5):601–615, 2006.

[27] David B Fogel. The advantages of evolutionary computation. In *BCEC*, pages 1–11, 1997.

[28] Melanie Mitchell. *An Introduction to Genetic Algorithms*. MIT press, 1998.

[29] Alden H Wright. Genetic algorithms for real parameter optimization. In *Foundations of genetic algorithms*, volume 1, pages 205–218. Elsevier, 1991.

[30] David M Deaven and Kai-Ming Ho. Molecular geometry optimization with a genetic algorithm. *Physical Review Letters*, 75(2):288, 1995.

[31] PDQ Adult Treatment Editorial Board. Colon cancer treatment (pdq®). In *PDQ Cancer Information Summaries [Internet]*. National Cancer Institute (US), 2020.

[32] Abhishek Bhandari, Melissa Woodhouse, and Samir Gupta. Colorectal cancer is a leading cause of cancer incidence and mortality among adults younger than 50 years in the usa: a seer-based analysis with comparison to other young-onset cancers. *Journal of Investigative Medicine*, 65(2):311–315, 2017.

[33] Rebecca Siegel, Carol DeSantis, and Ahmedin Jemal. Colorectal cancer statistics, 2014. *CA: A Cancer Journal for Clinicians*, 64(2):104–117, 2014.

[34] Justin Guinney, Rodrigo Dienstmann, Xin Wang, Aurelien De Reynies, Andreas Schlicker, Charlotte Soneson, Laetitia Marisa, Paul Roepman, Gift Nyamundanda, Paolo Angelino, et al. The consensus molecular subtypes of colorectal cancer. *Nature Medicine*, 21(11):1350, 2015.

[35] JR Jass. Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. *Histopathology*, 50(1):113–130, 2007.

[36] David Cunningham, Wendy Atkin, Heinz-Josef Lenz, Henry T Lynch, Bruce Minsky, Bernard Nordlinger, and Naureen Starling. Colorectal cancer. *The Lancet*, 375(9719):1030 – 1047, 2010.

[37] Rafael Lozano, Mohsen Naghavi, Kyle Foreman, Stephen Lim, Kenji Shibuya, Victor Aboyans, Jerry Abraham, Timothy Adair, Rakesh Aggarwal, Stephanie Y Ahn, Mohammad A AlMazroa, Miriam Alvarado, H Ross Anderson, Laurie M Anderson, Kathryn G Andrews, Charles Atkinson, Larry M Baddour, Suzanne Barker-Collo, David H Bartels, Michelle L Bell, Emelia J Benjamin, Derrick Bennett, Kavi Bhalla, Boris Bikbov, Aref Bin Abdulhak, Gretchen Birbeck, Fiona Blyth, Ian Bolliger, Soufiane Boufous, Chiara Bucello, Michael Burch, Peter Burney, Jonathan Carapetis, Honglei Chen, David Chou, Sumeet S Chugh, Luc E Coffeng, Steven D Colan, Samantha Colquhoun, K Ellicott Colson, John Condon, Myles D Connor, Leslie T Cooper, Matthew Corriere, Monica Cortinovis, Karen Courville de Vaccaro, William Couser, Benjamin C Cowie, Michael H Criqui, Marita Cross, Kaustubh C Dabhadkar, Nabila Dahodwala, Diego De Leo, Louisa Degenhardt, Allyne Delossantos, Julie Denenberg, Don C Des Jarlais, Samath D Dharmaratne, E Ray Dorsey, Tim Driscoll, Herbert Duber, Beth Ebel, Patricia J Erwin, Patricia Espindola, Majid Ezzati, Valery Feigin, Abraham D Flaxman, Mohammad H Forouzanfar, Francis Gerry R Fowkes, Richard Franklin, Marlene Fransen, Michael K Freeman, Sherine E Gabriel, Emmanuela Gakidou, Flavio Gaspari, Richard F Gillum, Diego Gonzalez-Medina, Yara A Halasa, Diana Haring, James E Harrison, Rasmus Havmoeller, Roderick J Hay, Bruno Hoen, Peter J Hotez, Damian

Hoy, Kathryn H Jacobsen, Spencer L James, Rashmi Jasrasaria, Sudha Jayaraman, Nicole Johns, Ganesan Karthikeyan, Nicholas Kassebaum, Andre Keren, Jon-Paul Khoo, Lisa Marie Knowlton, Olive Kobusingye, Adofo Koranteng, Rita Krishnamurthi, Michael Lipnick, Steven E Lipshultz, Summer Lockett Ohno, Jacqueline Mabweijano, Michael F MacIntyre, Leslie Mallinger, Lyn March, Guy B Marks, Robin Marks, Akira Matsumori, Richard Matzopoulos, Bongani M Mayosi, John H McAnulty, Mary M McDermott, John McGrath, Ziad A Memish, George A Mensah, Tony R Merriman, Catherine Michaud, Matthew Miller, Ted R Miller, Charles Mock, Ana Olga Mocumbi, Ali A Mokdad, Andrew Moran, Kim Mulholland, M Nathan Nair, Luigi Naldi, K M Venkat Narayan, Kiumarss Nasseri, Paul Norman, Martin O'Donnell, Saad B Omer, Katrina Ortblad, Richard Osborne, Doruk Ozgediz, Bishnu Pahari, Jeyaraj Durai Pandian, Andrea Panozo Rivero, Rogelio Perez Padilla, Fernando Perez-Ruiz, Norberto Perico, David Phillips, Kelsey Pierce, C Arden Pope, Esteban Porrini, Farshad Pourmalek, Murugesan Raju, Dharani Ranganathan, JÃŒrgen T Rehm, David B Rein, Guiseppe Remuzzi, Frederick P Rivara, Thomas Roberts, Felipe Rodriguez De LeÃ³n, Lisa C Rosenfeld, Lesley Rushton, Ralph L Sacco, Joshua A Salomon, Uchechukwu Sampson, Ella Sanman, David C Schwebel, Maria Segui-Gomez, Donald S Shepard, David Singh, Jessica Singleton, Karen Sliwa, Emma Smith, Andrew Steer, Jennifer A Taylor, Bernadette Thomas, Imad M Tleyjeh, Jeffrey A Towbin, Thomas Truelsen, Eduardo A Undurraga, N Venketasubramanian, Lakshmi Vijayakumar, Theo Vos, Gregory R Wagner, Mengru Wang, Wenzhi Wang, Kerrianne Watt, Martin A Weinstock, Robert Weintraub, James D Wilkinson, Anthony D Woolf, Sarah

Wulf, Pon-Hsiu Yeh, Paul Yip, Azadeh Zabetian, Zhi-Jie Zheng, Alan D Lopez, and Christopher JL Murray. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: A systematic analysis for the global burden of disease study 2010. *The Lancet*, 380(9859):2095 – 2128, 2012.

[38] Jacques Ferlay, Isabelle Soerjomataram, Rajesh Dikshit, Sultan Eser, Colin Mathers, Marise Rebelo, Donald Maxwell Parkin, David Forman, and Freddie Bray. Cancer incidence and mortality worldwide: Sources, methods and major patterns in globocan 2012. *International Journal of Cancer*, 136(5):E359–E386, 2015.

[39] Krapcho M Neyman N Aminou R Waldron W Altekruse SF Kosary CL Ruhl J Tatalovich Z Cho H Mariotto A Eisner MP Lewis DR Chen HS Feuer EJ Cronin KA (eds) Howlader N, Noone AM. Seer cancer statistics review, 1975-2009 (vintage 2009 populations),National Cancer Institute. Bethesda, MD, based on November 2011 SEER data submission, posted to the SEER web site, April 2012.

[40] Cristina Bosetti, Fabio Levi, Valentina Rosato, Paola Bertuccio, Franca Lucchini, Eva Negri, and Carlo La Vecchia. Recent trends in colorectal cancer mortality in europe. *International Journal of Cancer*, 129(1):180–191, 2011.

[41] Toshiaki Watanabe, Kei Muro, Yoichi Ajioka, Yojiro Hashiguchi, Yoshinori Ito, Yutaka Saito, Tetsuya Hamaguchi, Hideyuki Ishida, Megumi Ishiguro, Soichiro Ishihara, Yukihide Kanemitsu, Hiroshi Kawano, Yusuke Kinugasa, Norihiro Kokudo, Keiko Murofushi, Takako Nakajima, Shiro Oka, Yoshiharu Sakai, Aki-

hito Tsuji, Keisuke Uehara, Hideki Ueno, Kentaro Yamazaki, Masahiro Yoshida, Takayuki Yoshino, Narikazu Boku, Takahiro Fujimori, Michio Itabashi, Nobuo Koinuma, Takayuki Morita, Genichi Nishimura, Yuh Sakata, Yasuhiro Shimada, Keiichi Takahashi, Shinji Tanaka, Osamu Tsuruta, Toshiharu Yamaguchi, Naohiko Yamaguchi, Toshiaki Tanaka, Kenjiro Kotake, Kenichi Sugihara, and Japanese Society for Cancer of the Colon and Rectum. Japanese Society for Cancer of the Colon and Rectum (JSCCR) guidelines 2016 for the treatment of colorectal cancer. *International Journal of Clinical Oncology*, 23(1):1–34, Feb 2018.

[42] Melissa M Center, Ahmedin Jemal, Robert A Smith, and Elizabeth Ward. Worldwide variations in colorectal cancer. *CA: A Cancer Journal for Clinicians*, 59(6):366–378, 2009.

[43] AGE-ADJUSTED RATES and AGE-SPECIFIC RATES. SEER cancer statistics review 1975-2005. 2008.

[44] Peter Boyle and Michael JS Langman. ABC of colorectal cancer: Epidemiology. *BMJ*, 321(Suppl S6), 2000.

[45] Ivy Bazensky, Candice Shoobridge-Moran, and Linda H Yoder. Colorectal cancer: An overview of the epidemiology, risk factors symptoms, and screening guidelines. *Medsurg Nursing*, 16(1):46, 2007.

[46] Anna L Zisman, Angel Nickolov, Randall E Brand, Addi Gorchow, and Hemant K Roy. Associations between the age at diagnosis and location of col-

orectal cancer and the use of alcohol and tobacco: Implications for screening. *Archives of Internal Medicine*, 166(6):629–634, 2006.

[47] WH Tsong, WP Koh, JM Yuan, R Wang, CL Sun, and MC Yu. Cigarettes and alcohol in relation to colorectal cancer: The Singapore Chinese Health Study. *British Journal of Cancer*, 96(5):821–827, 2007.

[48] Elizabeth Half, Dani Bercovich, and Paul Rozen. Familial adenomatous polyposis. *Orphanet Journal of Rare Diseases*, 4(1):22, Oct 2009.

[49] Ian Tomlinson, Emily Webb, Luis Carvajal-Carmona, Peter Broderick, Zoe Kemp, Sarah Spain, Steven Penegar, Ian Chandler, Maggie Gorman, Wendy Wood, et al. A genome-wide association scan of tag snps identifies a susceptibility variant for colorectal cancer at 8q24. 21. *Nature Genetics*, 39(8):984, 2007.

[50] Mark M Pomerantz, Nasim Ahmadiyeh, Li Jia, Paula Herman, Michael P Verzi, Harshavardhan Doddapaneni, Christine A Beckwith, Jennifer A Chan, Adam Hills, Matt Davis, et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with myc in colorectal cancer. *Nature Genetics*, 41(8):882, 2009.

[51] Renata Hezova, Alena Kovarikova, Julie Bienertova-Vasku, Milana Sachlova, Martina Redova, Anna Vasku, Marek Svoboda, Lenka Radova, Igor Kiss, Rostislav Vyzula, et al. Evaluation of SNPs in mir-196-a2, mir-27a and mir-146a as risk factors of colorectal cancer. *World journal of Gastroenterology: WJG*, 18(22):2827, 2012.

[52] Henry T Lynch and Albert De la Chapelle. Hereditary colorectal cancer. *New England Journal of Medicine*, 348(10):919–932, 2003.

[53] CM Nutting. Cancer. Principles & practice of oncology (6th edn). Ed. by VT Devita Jr, S Hellman and SA Rosenburg, pp. lxxii+ 3235, 2001 (Lippincott Williams & Wilkins, Philadelphia, PA), ISBN 0-781-72229-2, 2002.

[54] Jeannette Jackson-Thompson, Faruque Ahmed, Robert R German, Sue-Min Lai, and Carol Friedman. Descriptive epidemiology of colorectal cancer in the united states, 1998–2001. *Cancer*, 107(S5):1103–1111, 2006.

[55] D Max Parkin, Freddie Bray, J Ferlay, and Paola Pisani. Global cancer statistics, 2002. *CA: A Cancer Journal for Clinicians*, 55(2):74–108, 2005.

[56] Ketan Thanki, Michael Edward Nicholls, Aakash Gajjar, Anthony J Senagore, Suimin Qiu, Csaba Szabo, Mark R Hellmich, and Celia Chao. Consensus molecular subtypes of colorectal cancer and their clinical implications. *International Biological and Biomedical Journal*, 3(3):105–111, 2017.

[57] Francis S Collins, Michael Morgan, and Aristides Patrinos. The human genome project: Lessons from large-scale biology. *Science*, 300(5617):286–290, 2003.

[58] Mark D Adams, Jenny M Kelley, Jeannine D Gocayne, Mark Dubnick, Mihael H Polymeropoulos, Hong Xiao, Carl R Merril, Andrew Wu, Bjorn Olde, Ruben F Moreno, et al. Complementary dna sequencing: expressed sequence tags and human genome project. *Science*, 252(5013):1651–1656, 1991.

[59] Francis S Collins and Victor A McKusick. Implications of the human genome project for medical science. *JAMA*, 285(5):540–544, 2001.

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

[61] William YS Wang, Bryan J Barratt, David G Clayton, and John A Todd. Genome-wide association studies: Theoretical and practical concerns. *Nature Reviews Genetics*, 6(2):109–118, 2005.

[62] Andrea S Foulkes. *Applied Statistical Genetics with R*. Springer, 2009.

[63] Mark I McCarthy, Gonçalo R Abecasis, Lon R Cardon, David B Goldstein, Julian Little, John PA Ioannidis, and Joel N Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369, 2008.

[64] Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five years of GWAS discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.

[65] Arthur Korte and Ashley Farlow. The advantages and limitations of trait analysis with GWAS: A review. *Plant Methods*, 9(1):1–9, 2013.

[66] Lucia A Hindorff, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri 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.

[67] Wellcome Trust Case Control 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.

[68] Rita M Cantor, Kenneth Lange, and Janet S Sinsheimer. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *The American Journal of Human Genetics*, 86(1):6–22, 2010.

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

[70] McGraw Hill Tom Mitchell. *Machine Learning*. McGraw-Hill Science/Engineering/Math; (March 1, 1997), 1997.

[71] Pedro Larranaga, Borja Calvo, Roberto Santana, Concha Bielza, Josu Galdiano, Inaki Inza, José A Lozano, Rubén Armananzas, Guzmán Santafé, Aritz Pérez, et al. Machine learning in bioinformatics. *Briefings in Bioinformatics*, 7(1):86–112, 2006.

[72] Maxwell W Libbrecht and William Stafford Noble. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321–332, 2015.

[73] Silke Szymczak, Joanna M Biernacka, Heather J Cordell, Oscar González-Recio, Inke R König, Heping Zhang, and Yan V Sun. Machine learning in genome-wide association studies. *Genetic Epidemiology*, 33(S1):S51–S57, 2009.

[74] Ting Hu, Nicholas A Sinnott-Armstrong, Jeff W Kiralis, Angeline S Andrew, Margaret R Karagas, and Jason H Moore. Characterizing genetic interactions in human disease association studies using statistical epistasis networks. *BMC Bioinformatics*, 12(1):1–13, 2011.

[75] Peng Yang, Xiao-Li Li, Jian-Ping Mei, Chee-Keong Kwoh, and See-Kiong Ng. Positive-unlabeled learning for disease gene identification. *Bioinformatics*, 28(20):2640–2647, 2012.

[76] Hong-Woo Chun, Yoshimasa Tsuruoka, Jin-Dong Kim, Rie Shiba, Naoki Nagata, Teruyoshi Hishiki, and Jun'ichi Tsujii. Extraction of gene-disease relations from medline using domain dictionaries and machine learning. In *Biocomputing 2006*, pages 4–15. World Scientific, 2006.

[77] Malgorzata Maciukiewicz, Victoria S Marshe, Anne-Christin Hauschild, Jane A Foster, Susan Rotzinger, James L Kennedy, Sidney H Kennedy, Daniel J Müller, and Joseph Geraci. GWAS-based machine learning approach to predict duloxetine response in major depressive disorder. *Journal of Psychiatric Research*, 99:62–68, 2018.

[78] Bing Han, Meeyoung Park, and Xue-wen Chen. A Markov blanket-based method for detecting causal SNPs in GWAS. In *BMC Bioinformatics*, volume 11, page S5. Springer, 2010.

[79] Jaume Bacardit and Xavier Llorà. Large-scale data mining using genetics-based machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):37–61, 2013.

[80] Avrim L Blum and Pat Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2):245–271, 1997.

[81] Norshafarina Omar, Fatimatufaridah Jusoh, R Ibrahim, and MS Othman. Review of feature selection for solving classification problems. *Journal of Information System Research and Innovation*, 3:64–70, 2013.

[82] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*, volume 112. Springer, 2013.

[83] Jeovany Martínez-Mesa, David Alejandro González-Chica, Rodrigo Pereira Duquia, Renan Rangel Bonamigo, and João Luiz Bastos. Sampling: How to select participants in my research study? *Anais Brasileiros de Dermatologia*, 91(3):326–330, 2016.

[84] Rui Xia, Chengqing Zong, Xuelei Hu, and Erik Cambria. Feature ensemble plus sample selection: domain adaptation for sentiment classification. *IEEE Intelligent Systems*, 28(3):10–18, 2013.

[85] Jason H Moore, Folkert W Asselbergs, and Scott M Williams. Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, 26(4):445–455, 2010.

[86] Nada MA Al-Salami. Evolutionary algorithm definition. *American Journal of Engineering and Applied Sciences*, 2(4):789–795, 2009.

[87] P. A. Vikhar. Evolutionary algorithms: A critical review and its future prospects. In *2016 International Conference on Global Trends in Signal Process-*

ing, Information Computing and Communication (ICGTSPICC), pages 261–265, 2016.

[88] John Henry Holland et al. *Adaptation in Natural and Artificial systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. MIT press, 1992.

[89] Xin Yao. Global optimisation by evolutionary algorithms. In *Proceedings of IEEE International Symposium on Parallel Algorithms Architecture Synthesis*, pages 282–291. IEEE, 1997.

[90] Pradnya Vikhar. Evolutionary algorithm: A classical search and optimization technique. *International Journal of Pure and Applied Research in Engineering and Technology*, 4(9):758–766, 2016.

[91] AJ Keane. The design of a satellite beam with enhanced vibration performance using genetic algorithm techniques. *Journal of the Acoustical Society of America*, 99(4):2599–2603, 1996.

[92] Sonia Schulenburg and Peter Ross. Strength and money: An LCS approach to increasing returns. In *International Workshop on Learning Classifier Systems*, pages 114–137. Springer, 2000.

[93] Larry J Eshelman. Preventing premature convergence in genetic algorithms by preventing incest. In *Proc. of the 4th. Conf. on GA, July, 1991*, pages 115–122, 1991.

[94] Agoston E Eiben and James E Smith. *Introduction to Evolutionary Computing*. Springer, 2015.

[95] Agoston E Eiben, Emile HL Aarts, and Kees M Van Hee. Global convergence of genetic algorithms: A markov chain analysis. In *International Conference on Parallel Problem Solving from Nature*, pages 3–12. Springer, 1990.

[96] David E Goldberg and John Henry Holland. Genetic algorithms and machine learning. 1988.

[97] Darrell Whitley. A genetic algorithm tutorial. *Statistics and computing*, 4(2):65–85, 1994.

[98] Manoj Kumar, Mohamed Husain, Naveen Upreti, and Deepti Gupta. Genetic algorithm: Review and application. *Available at SSRN 3529843*, 2010.

[99] Daniel S Weile and Eric MichieFylssen. Genetic algorithm optimization applied to electromagnetics: A review. *IEEE Transactions on Antennas and Propagation*, 45(3):343–353, 1997.

[100] Minglun Gong and Yee-Hong Yang. Multi-resolution stereo matching using genetic algorithm. In *Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001)*, pages 21–29. IEEE, 2001.

[101] Praveen Ranjan Srivastava and Tai-hoon Kim. Application of genetic algorithm in software testing. *International Journal of software Engineering and its Applications*, 3(4):87–96, 2009.

[102] Hxugo Kubiny. Variable selection in QSAR studies. I. An evolutionary algorithm. *Quantitative Structure-Activity Relationships*, 13(3):285–294, 1994.

[103] Kay Chen Tan, Eu Jin Teoh, Q Yu, and KC Goh. A hybrid evolutionary algorithm for attribute selection in data mining. *Expert Systems with Applications*, 36(4):8616–8630, 2009.

[104] Riyaz Sikora and Selwyn Piramuthu. Framework for efficient feature selection in genetic algorithm based data mining. *European Journal of Operational Research*, 180(2):723–737, 2007.

[105] W Daniel Hillis. Co-evolving parasites improve simulated evolution as an optimization procedure. *Physica D: Nonlinear Phenomena*, 42(1-3):228–234, 1990.

[106] Robert Hinterding. Self-adaptation using multi-chromosomes. In *Proceedings of 1997 IEEE International Conference on Evolutionary Computation (ICEC'97)*, pages 87–91. IEEE, 1997.

[107] James Alfred Walker, Julian Francis Miller, and Rachel Cavill. A multi-chromosome approach to standard and embedded cartesian genetic programming. In *Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation*, pages 903–910, 2006.

[108] Rachel Cavill, Hector C Keun, Elaine Holmes, John C Lindon, Jeremy K Nicholson, and Timothy MD Ebbels. Genetic algorithms for simultaneous variable and sample selection in metabonomics. *Bioinformatics*, 25(1):112–118, 2008.

[109] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A.R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I.W. de Bakker, Mark J. Daly, and Pak C. Sham. Plink: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559 – 575, 2007.

[110] Carl A Anderson, Fredrik H Pettersson, Geraldine M Clarke, Lon R Cardon, Andrew P Morris, and Krina T Zondervan. Data quality control in genetic case-control association studies. *Nature protocols*, 5(9):1564–1573, 2010.

[111] Cathy C Laurie, Kimberly F Doheny, Daniel B Mirel, Elizabeth W Pugh, Laura J Bierut, Tushar Bhangale, Frederick Boehm, Neil E Caporaso, Marilyn C Cornelis, Howard J Edenberg, et al. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic Epidemiology*, 34(6):591–602, 2010.

[112] Stephen Turner, Loren L Armstrong, Yuki Bradford, Christopher S Carlson, Dana C Crawford, Andrew T Crenshaw, Mariza de Andrade, Kimberly F Doheny, Jonathan L Haines, Geoffrey Hayes, et al. Quality control procedures for genome-wide association studies. *Current Protocols in Human Genetics*, 68(1):1–19, 2011.

[113] Jacqueline K Wittke-Thompson, Anna Pluzhnikov, and Nancy J Cox. Rational inferences about departures from Hardy-Weinberg equilibrium. *The American Journal of Human Genetics*, 76(6):967–986, 2005.

[114] SG Larmer, M Sargolzaei, and FS Schenkel. Extent of linkage disequilibrium, consistency of gametic phase, and imputation accuracy within and across canadian dairy breeds. *Journal of Dairy Science*, 97(5):3128–3141, 2014.

[115] GR Wiggans, TS Sonstegard, PM VanRaden, LK Matukumalli, RD Schnabel, JF Taylor, FS Schenkel, and CP Van Tassell. Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the united states and canada. *Journal of Dairy Science*, 92(7):3431–3436, 2009.

[116] Kenji Kira and Larry A Rendell. A practical approach to feature selection. In *Machine Learning Proceedings 1992*, pages 249–256. Elsevier, 1992.

[117] Igor Kononenko, Edvard Šimec, and Marko Robnik-Šikonja. Overcoming the myopia of inductive learning algorithms with RELIEFF. *Applied Intelligence*, 7(1):39–55, 1997.

[118] Casey S Greene, Nadia M Penrod, Jeff Kiralis, and Jason H Moore. Spatially uniform ReliefF (SURF) for computationally-efficient filtering of gene-gene interactions. *BioData mining*, 2(1):5, 2009.

[119] Jason H Moore and Bill 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.

[120] Ryan J Urbanowicz, Melissa Meeker, William La Cava, Randal S Olson, and Jason H Moore. Relief-based feature selection: Introduction and review. *Journal of Biomedical Informatics*, 85:189–203, 2018.

[121] Somayeh Kafaie, Yuanzhu Chen, and Ting Hu. A network approach to prioritizing susceptibility genes for genome-wide association studies. *Genetic Epidemiology*, 43(5):477–491, 2019.

[122] Naomi S Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[123] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth Annual Workshop on Computational Learning Theory*, pages 144–152, 1992.

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

[125] Félix-Antoine Fortin, François-Michel De Rainville, Marc-André Gardner Gardner, Marc Parizeau, and Christian Gagné. DEAP: Evolutionary algorithms made easy. *The Journal of Machine Learning Research*, 13(1):2171–2175, 2012.

[126] 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.

[127] Yoichiro Kakugawa, Tadashi Wada, Kazunori Yamaguchi, Hideaki Yamanami, Kiyoaki Ouchi, Ikuro Sato, and Taeko Miyagi. Up-regulation of plasma membrane-associated ganglioside sialidase (neu3) in human colon cancer and its involvement in apoptosis suppression. *Proceedings of the National Academy of Sciences*, 99(16):10718–10723, 2002.

[128] Kohta Takahashi, Masahiro Hosono, Ikuro Sato, Keiko Hata, Tadashi Wada, Kazunori Yamaguchi, Kazuo Nitta, Hiroshi Shima, and Taeko Miyagi. Sialidase neu 3 contributes neoplastic potential on colon cancer cells as a key modulator of gangliosides by regulating w nt signaling. *International Journal of Cancer*, 137(7):1560–1573, 2015.

[129] Kazuhiro Shiozaki, Kazunori Yamaguchi, Ikuro Sato, and Taeko Miyagi. Plasma membrane-associated sialidase (neu3) promotes formation of colonic aberrant crypt foci in azoxymethane-treated transgenic mice. *Cancer Science*, 100(4):588–594, 2009.

[130] LI Wang, Dequan Liu, Xingrao Wu, Yueqin Zeng, Lan Li, Yu Hou, Wenhui Li, and Zhijie Liu. Long non-coding rna (LncRNA) rmst in triple-negative breast cancer (TNBC): Expression analysis and biological roles research. *Journal of Cellular Physiology*, 233(10):6603–6612, 2018.

[131] Remco Nagel, Carlos le Sage, Begona Diosdado, Maike van der Waal, Joachim AF Oude Vrielink, Anne Bolijn, Gerrit A Meijer, and Reuven Agami. Regulation of the adenomatous polyposis coli gene by the mir-135 family in colorectal cancer. *Cancer Research*, 68(14):5795–5802, 2008.

[132] Yiwen Fang and Melissa J Fullwood. Roles, functions, and mechanisms of long non-coding RNAs in cancer. *Genomics, Proteomics & Bioinformatics*, 14(1):42–54, 2016.

[133] Jirong Long, Qiuyin Cai, Hyuna Sung, Jiajun Shi, Ben Zhang, Ji-Yeob Choi, Wanqing Wen, Ryan J Delahanty, Wei Lu, Yu-Tang Gao, et al. Genome-wide

association study in east asians identifies novel susceptibility loci for breast cancer. *PLoS Genet*, 8(2):e1002532, 2012.